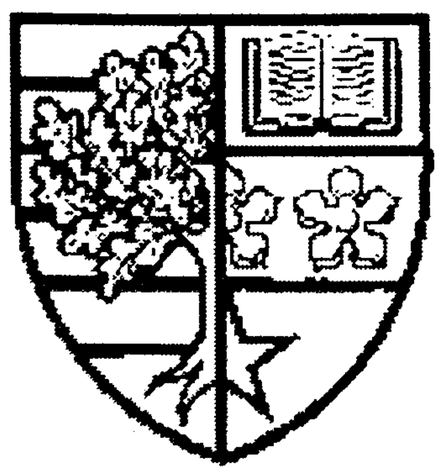


Identifying Functionally and Topologically Cohesive Modules in Protein Interaction Networks

Zelmina Lubovac

Submitted for the degree of Doctor of Philosophy in Computer Science

Heriot-Watt University

School of Mathematical and Computer Sciences

April 2008

Abstract

Systems biology offers a holistic perspective where individual proteins are viewed as elements in a network of protein-protein interactions, in which the proteins have contextual functions within functional modules. In order to facilitate the identification and analysis of such modules, we here propose novel approaches that combine knowledge, in terms of Gene Ontology annotation with network topology information. The majority of previous methods for identifying modules in protein interaction networks are based solely on analysing topological features of the networks. In contrast, we propose the use of integrated functionally informed clustering coefficients to identify biologically plausible network modules. The main part of the thesis is focused on the method named SWEMODE (Semantic WEights for MODule Elucidation), which uses a weighted clustering coefficient to identify network functional modules. We demonstrate that the proposed methods are able to identify the key functional modules in protein interaction networks. We also investigate the functional and topological features of the proteins that are involved in multiple modules, as well as their role in the interconnectivity between modules. The results indicate that the majority of, so called multi-modular proteins are involved in the assembly and arrangement of cell structures, such as the cell wall and cell envelope.

Acknowledgements

First and foremost, I would like to deeply thank my supervisors, Björn Olsson and David Corne for their enormous support and guidance. Björn, my supervisor at the University of Skövde, has led me through both my MSc and PhD projects, and without his valuable insights and generous help with all aspects of my thesis work, this thesis would not have been possible. David's experienced pieces of advice were invaluable for the completion of this thesis. Thanks must also go to Ajit Narayanan, my former supervisor, for his enthusiasm that was one of the reasons for me to pursue this area of research.

Special thanks to the examiners of my thesis for taking the time to review my work.

Thanks to all colleagues, especially Jonas Gamalielsson, for providing technical support, being a critical eye, and making his famous jokes. Thanks to Angelica Lindlöf, for all valuable discussions, and for being a great colleague and an even better friend.

I am very grateful to my parents, the main reason that I am here, for their endless love and encouragement. I also thank my sister who used to sabotage my work on the PhD thesis by stealing my weekend hours. Thanks should also go to my parents in low for standing by me during my PhD years.

My PhD years will always be the most special part of my life, partly thanks to two little persons that came into my life. The first being my daughter Melina, who came into the world in the middle of my PhD studies, and taught me to be more efficient, focused and structured. The second being my baby Alma who taught me that it is very difficult to plan the exact time of delivery, when it comes to babies and theses.

I am deeply grateful to my fiancé Amir, who was always there for me through thick and thin.

Finally, thanks to all friends and relatives for putting up with me through some tough periods.

ACADEMIC REGISTRY
Research Thesis Submission

Name:	Zelmina Lubovac		
School/PGI:	School of Mathematical and Computer Sciences		
Version: *(i.e. First, Resubmission, Final)*	Final	Degree Sought:	PhD

Declaration

In accordance with the appropriate regulations I hereby submit my thesis and I declare that:

1) the thesis embodies the results of my own work and has been composed by myself
2) where appropriate, I have made acknowledgement of the work of others and have made reference to work carried out in collaboration with other persons
3) the thesis is the correct version of the thesis for submission and is the same version as any electronic versions submitted*.
4) my thesis for the award referred to, deposited in the Heriot-Watt University Library, should be made available for loan or photocopying and be available via the Institutional Repository, subject to such conditions as the Librarian may require
5) I understand that as a student of the University I am required to abide by the Regulations of the University and to conform to its discipline.

* *Please note that it is the responsibility of the candidate to ensure that the correct version of the thesis is submitted.*

Signature of Candidate:	Zelmina Lubovac	Date:	27/5 - 2008

Submission

Submitted By *(name in capitals)*:	ZELMINA LUBOVAC
Signature of Individual Submitting:	Zelmina Lubovac / Prelim Scan
Date Submitted:	27/5 - 2008 29/5/08

For Completion in Academic Registry

Received in the Academic Registry by *(name in capitals)*:	J TOUGH		
Method of Submission *(Handed in to Academic Registry; posted through internal/external mail)*:	HANDED IN		
E-thesis Submitted	NO		
Signature:	J TS	Date:	29.5.08 .

Dissemination

Zelmina Lubovac, David Corne, Jonas Gamalielsson, and Björn Olsson (2007): Weighted Cohesiveness for Identification of Functional Modules and their Interconnectivity. In: Hochreiter, S. and Wagner, R. (eds.), *Proceedings of Bioinformatics Research and Development* BIRD 2007, Berlin, Germany, March 12-14. LNBI 4414: 185-198, Springer Verlag.

Zelmina Lubovac, Jonas Gamalielsson, and Björn Olsson (2006): Combining functional and topological properties to identify core modules in protein interaction networks. *PROTEINS: Structure, Function and Bioinformatics*, 64(4):948-959.

Zelmina Lubovac, Björn Olsson, and Jonas Gamalielsson (2006): Weighted Clustering Coefficient for Identifying Modular Formations in Protein-Protein Interaction Networks. *Proceedings of the Third International Conference on Bioinformatics and Computational and Systems Biology* BCSB 2006, Prague, Czech Republic, August 25-27, 122-127.

Zelmina Lubovac, Björn Olsson, and Jonas Gamalielsson (2005): Combining topological characteristics and domain knowledge reveals functional modules in protein interaction networks. In Sagot, M-F. and Guimaraes, K.S., *Proceedings of the Second Conference on Algorithms and computational methods for Biochemical and evolutionary networks* CompBioNets 2005, Lyon, France, December 5-7, 93-106. College Publications.

Zelmina Lubovac, Jonas Gamalielsson, Björn Olsson, and Angelica Lindlöf (2005): Exploring protein networks with a semantic similarity measure across Gene Ontology. *Proceedings of the Sixth International Symposium on Computational Biology and Genome Informatics* CBGI 2005, Salt Lake City, USA, July 21-26, 1203-1208.

Table of Contents

List of Tables

List of Figures

Index of Acronyms

CYGD: the Comprehensive Yeast Genome Database

DAG: Direct Acyclic Graph

DIP: Database of Interacting Proteins

DFS: Depth-First Search

HMS-PCI: High-throughput Mass Spectrometric Protein Complex Identification

KEGG: Kyoto Encyclopedia of Genes and Genomes

MAPK: Mitogen-Activated Protein Kinase

MIPS: the Munich Information center for Protein Sequences

MS: Mass Spectrometry

NWP: Node Weight Percentage

PCP: Protein-Complex Purification

PIN: Protein Interaction Network

SWEMODE: Semantic WEights for MODule Elucidation

TAP: Tandem Affinity Purification

Y2H: Yeast Two-Hybrid

Chapter 1

Introduction

One of the challenges that systems biology is facing consists of explaining biological organisation in the light of the existence of modules in networks (Han, et al., 2004; Pereira-Leal, et al., 2004; Petti and Church, 2005; Rives and Galitski, 2003). A proposal that cellular function is carried out by modules, (Hartwell, et al., 1999) has fired a "modular era" of systems biology in which the focus has been on studying modularity at different levels of cellular organisation. A series of studies attempting to reveal the modules in cellular networks, ranging from metabolic (Ravasz, et al., 2002), to protein networks (Spirin and Mirny, 2003; Yook, et al., 2004), support the proposal that modular architecture is one of the principles underlying biological organisation.

The term "module", as understood in molecular biology, was originally defined as a discrete unit with a function that is separable from those of other modules (Hartwell, et al., 1999). The separability may originate from, for example, cellular localization of specific protein interactions. Furthermore, modularity involves groups of elements that work in a co-operative fashion to achieve some defined function. In a general network representation, modules appear as highly interconnected groups of nodes (Barabasi and Oltvai, 2004). Protein complexes constitute one example of a type of module, since the proteins within a complex interact functionally and physically to form a robust unit, which in its turn carries out some biological function (Yook, et al., 2004). Many other kinds of modules, however, consist of proteins that do not interact physically and directly with each other, but nevertheless are involved in carrying out the same function.

Understanding protein interactions and studying networks of these interactions provide valuable insights into the complexity and the structural organisation of cells. This

understanding may help to uncover the generic organising principles of cellular networks. Several studies show that modularity is one such principle (Gavin, et al., 2006; Han, et al., 2004; Qi and Ge, 2006; Rives and Galitski, 2003).

Although numerous experimental methods are available for high-throughput identification of protein-protein interactions, the most widely accepted are the yeast two-hybrid (Y2H) system and a combination of protein-complex purification and mass spectrometry (MS) (Fields and Bartel, 2001; Mann, et al., 2001). Protein interactions identified on a genome-wide scale are commonly represented as protein interaction networks (PINs). Such networks are graphs with nodes corresponding to proteins and edges corresponding to interactions. An example of a PIN constructed in this way can be seen in Figure 1 below. The network represents 1870 proteins, connected by 2240 direct physical interactions.

Figure 1: Example of yeast PIN (Jeong, et al., 2001). The node colour represents the phenotypic effect of removing the corresponding proteins (red = lethal, green = non lethal, orange = slow growth, yellow = unknown)

Such networks, although not capturing all aspects of the true dynamics within the cell, allow for the analysis of certain topological and functional properties that may help uncover the organising principles that govern their formation and evolution.

Many biological networks, including PINs and metabolic networks, have a so-called scale-free topology (Barabasi and Albert, 1999) meaning that the number of nodes with a given degree follows a power law. Networks with power law degree distributions are highly non-uniform: the majority of nodes have few links, whereas there are a few nodes with very large numbers of links, called hubs (see example network in Figure 1 on page 2). For example, a spindle-pole body corresponds to a group of protein complexes that form a hub for the attachment and organisation of microtubules (Rives and Galitski, 2003). Furthermore, the coexistence of the scale-free property and a high clustering coefficient in cellular networks is a signature of so called hierarchical modularity, indicating that topological modules are not independent units, but combine to form a hierarchical network (Ravasz and Barabasi, 2003; Ravasz, et al., 2002). The clustering coefficient of a node (protein) is a measure of the interconnectivity of the neighbours of a node. Simply stated, it measures the density of triangles, i.e. triplets of interconnected proteins, formed in the immediate neighbourhood of the node.

Although topology-based measures, such as clustering coefficient, are very important, and may be used for identifying functional modules, there are some reasons why we should consider functional knowledge as well when deriving modules. High-throughput protein interaction data that is often used to identify modules is very noisy (Titz, et al., 2004). Technologies such as Y2H often result in many false positives that may cause misleading conclusions in the analysis. A possible approach to decrease the number of false interactions may be to focus on the "high confidence" data sets, where all interactions have been confirmed by several experiments. However, this approach discards the majority of the existing interactions. A better approach could be achieved by incorporating the available knowledge associated with available interactions into the analysis. This has also been pointed out in previous studies that focus on deriving protein complexes by using topological information. In (Przulj, et al., 2004), it has been observed that the increasing size of PINs (by including medium and low confidence interactions) has resulted in a decreasing number of highly connected sub-graphs or clusters which may correspond to protein complexes. As Przulj, et al., (2004) state, this is probably due to the increasing noise in the data, and a possible solution to this problem may be the integration of PINs with annotation or gene expression data.

According to Bader and Hogue (2003) "more accurate data mining algorithms and systems models could be constructed to understand and predict interactions, complexes and pathways by taking into account more existing biological knowledge."

Another reason for considering knowledge represented in annotations a valuable complement to topological characteristics is encompassed in the concept of functional modules themselves. A functional module consists of proteins that cooperate towards achieving a particular function or participate in similar processes. Hence, considering annotation that describes molecular functions and biological processes should enrich the protein-protein interactions.

An ultimate description of cellular networks would require considering both the strength and the temporal aspects of the interactions (Barabasi and Oltvai, 2004). However, in spite of recent advances, molecular network biology is still in its infancy, and future progress will require the development of highly sensitive tools for quantifying the concentrations and interactions at high resolution in both time and space.

In summary, our post-genomic view is expanding the role of the protein into an element in a network of protein-protein interactions, in which it has contextual functions within functional modules (Eisenberg, et al., 2000; Jeong, et al., 2001). This raises several questions, which are addressed in current research in systems biology. How do modules interact to achieve a certain functionality (Han, et al., 2004; Rives and Galitski, 2003)? How can we evaluate the biological relevance of modules (Pereira-Leal, et al., 2004; Poyatos and Hurst, 2004)? Answering those questions may facilitate our understanding of the relationships between structure, function and regulation of molecular networks, which is one of the important aims of systems biology (Qi and Ge, 2006; Stelling, et al., 2002).

To contribute to this goal, we focus our attention on integrating the topology, i.e., various structural properties of the networks, with the functional knowledge represented in protein annotations. The methods we developed in this work are able to generate functional modules that may serve as predictive models for hypothesis generation.

One problem associated with module-identifying procedures is the lack of objective criteria for what should be considered a module. The proposed framework for identifying modules described in Chapters 6-8, is based on defining the module as a dense region of the PIN, which contains functionally related proteins. In previous work, an algorithm for Molecular COmplex DEtection (MCODE), based on a local network

density function named *core-clustering coefficient*, has been proposed (Bader and Hogue, 2003). Also, several methods of network clustering have been applied to reveal modular organisation in PINs (Pereira-Leal, et al., 2004; Poyatos and Hurst, 2004; Rives and Galitski, 2003). However, those methods have been focused on topological properties of the network. In contrast, this thesis investigates the characterisation of modules by using a combination of *topological* and *semantic* information. We define measures of *semantic cohesiveness* for modules, and these measures are calculated using both topological properties of the network, and information obtained from the Gene Ontology (GO) annotations of the proteins involved. In this work, we use novel measures, *semantically weighted clustering coefficients*, which integrate topological characteristics of the network with semantic similarity based on the functional similarity between interacting proteins. We also combine clustering, based on semantic similarity profiles with mutual neighbours-based clustering, to obtain modular structures based on both aspects.

A module-identifying algorithm, SWEMODE (Semantic WEights for MODule Elucidation), based on semantic cohesiveness, is developed here to test and evaluate the proposed metrics. Another method, based on a novel combination of semantic similarity profiles and mutual neighbours profiles is proposed and applied to known modular networks, to test if it is able to recover known modules, and to evaluate the interconnectedness between modules. The proposed methods take advantage of three aspects of functional annotation encoded in GO, namely molecular function, biological process and cellular component, and combine these with topological properties of the protein network. An additional aim of this work is to investigate if combining functional and topological information is useful for describing global properties of the protein network.

1.1 Context of the thesis

At a high level in the complexity pyramid of life (Oltvai and Barabasi, 2002) (see Figure 3 on page 17), protein complexes and proteins may interact weakly and transiently but may also be cohesive and connect strongly with preferred partners to form modules that serve distinct functions. Modules are often seen as an abstraction of complexes. However, one important distinction between complexes and functional modules is that complexes correspond to groups of proteins that interact with each other at the same time and place, forming a single multi-molecular machine. Examples of

protein complexes include the anaphase-promoting complex, origin recognition complex, protein export and transport complexes, etc. Functional modules, in contrast, do not require physical interaction between all the components at the same point in time, but rather consist of proteins that participate in a particular *cellular process* while binding to each other at different times and places, such as different conditions or phases of the cell cycle (Pereira-Leal, et al., 2006). Examples of functional modules include the yeast pheromone response pathway, MAPK signalling cascades, etc. Furthermore, not all functional modules require a physical interaction between components (Pereira-Leal, et al., 2006). In summary, a functional module may be conceptualised as a process (Schlosser, 2004) which does not necessarily correspond to a structure defined in time and in space, like a protein complex.

Consequently, integrated approaches that combine topology information with knowledge that describes how gene products behave in a cellular context, such as taking part in certain biological process or building up an anatomical structure, should be useful for providing new insights about functional modules. The GO Consortium (2001) provides three separate ontologies – molecular function, biological process and cellular component – to describe the important attributes of gene products that we seek to integrate with topology, to identify functional modules.

1.2 Questions investigated

Functional modules correspond to dense sub-graphs that contain proteins that participate in the same biological processes or act together to perform a distinct biological function. The aim of the work reported in this thesis is to develop integrated methods that combine semantic and topological information to identify such modules in protein interaction networks. We also aim to investigate their properties and potential advantages in comparison to the corresponding methods which do not take into account semantic information represented in biological annotations.

In order to reach this aim, we designed, implemented and tested novel methods that integrate semantic information about proteins, such as their involvement in biological processes, with topological measures of protein networks, such as clustering coefficients, for the purpose of identifying functional modules in protein interaction networks.

One way to capture knowledge about protein interactions is to use the information stored in a structured vocabulary that covers various knowledge-based aspects of molecular biology. GO is one of the most important ontologies within the bioinformatics community which can be used to describe gene and gene product attributes in any organism. It is considered as a *de facto* standard in molecular biology for the annotation of proteins. We use this ontology, because it provides a collection of well-defined biological terms, spanning molecular function, biological processes and cellular components. Each of these sub-ontologies reflects a certain functional aspect, and is therefore considered suitable for the purpose of module identification

Various measures based on information content have been applied to measure the similarity between GO terms. There are three semantic similarity measures that have been most frequently applied to GO terms. Since it is not clear which of these measures is the most advantageous for the purpose of this work, the following question was investigated:

Question 1: What are the properties of semantic similarity measures, and what is the most appropriate way to use semantic similarity measures to calculate weights, in terms of similarity between proteins?

Once we have chosen a suitable semantic similarity measure for calculating the weights, the next step is to choose a topological measure that is to be combined with the measure based on semantic information. It is known that PINs are characterised by a high degree of clustering, which is also one of the signatures for potential modularity in scale-free networks. This means that the probability that the neighbours of a node are also neighbours of each other in such networks is higher than in random networks. Consequently, for any edge *ij* between nodes *i* and *j* from a scale-free modular network, the neighbour of *i* is more likely to have an edge to *j*, than would be the case in a random network. As pointed out by Goldberg and Roth (2003), such "mutual neighbours" of the two endpoints serve to corroborate the edge. They defined a mutual clustering coefficient for a pair of nodes to give a measure of such corroboration. This measure does not treat each pair of proteins individually, but in relation to all other proteins in the network. Furthermore, in the method that we describe in Chapter 5, we assume that proteins belonging to the same module share neighbours to a great extent and are functionally similar. Hence, combining this topological measure with the measure based on semantic information (that describes functions and processes that

proteins are involved in) could be advantageous for module identification, and we therefore investigated the following research question:

Question 2: Can we gain any additional knowledge by extracting a modular structure based on both functional knowledge stored in the annotations and topological information (mutual neighbours profiles), compared to the strategy of extracting the corresponding modules based solely on topological properties?

It is known that large-scale protein interaction data, although proven useful for identifying protein complexes, suffers from a high error rate, in terms of false positive interactions. The lesson learned from Y2H is that this technique is more reliable for finding transient interactions, whereas more stable interactions will not always be detected with Y2H (Titz, et al., 2004). The presence of noisy edges makes it difficult to define a quality measure based purely on topology, such as when using a density measure. Therefore, we propose the concept of module semantic cohesiveness, based on both topological and semantic information, to describe clusters of proteins that are not only densely connected but also perform similar functions or participate in the same biological processes. For that purpose, we use the terms stored in biological ontologies, where different biological aspects may be covered by different sub-ontologies. The following questions were investigated in this context:

Question 3: What would be an appropriate way to measure topological and semantic cohesiveness?

Question 4: Which of the three aspects that are covered in GO is most appropriate for identifying modules in PINs? Is it beneficial for this approach to combine several aspects?

One of the graph theoretic properties that have recently been applied in biological contexts is the concept of *k*-cores of a graph. The concept of *k*-cores has been applied for finding densely connected protein complexes (Bader and Hogue, 2003), identifying important and evolutionarily conserved proteins (Wuchty and Almaas, 2005), visualisation of biological networks (Adamcsek, et al., 2006), etc. Our ambition is to further investigate the properties of the semantically weighted clustering coefficient, proposed as one of the scoring functions for module identification, by exploring it on the basis of alternative topological properties. Using *k*-cores of the graph may help to discern the highly interconnected sub-graphs of proteins while dismissing less connected proteins (for example singly linked proteins). This property is concerned directly with topological cohesiveness, which in turn is an important aspect of the way

we define modules. Therefore, it is an interesting candidate for an alternative topological property to be combined with the proposed weighted measure. The following question was investigated:

Question 5: Given that we are investigating measures of semantic cohesiveness based on both topological and semantic properties, what difference does it make if we base the topology part of the measure on the *k*-core sub-graph of the PIN, rather than the original PIN?

To generate functional modules as functionally and structurally cohesive structures in PINs is an important step towards reaching the top of the life's complexity pyramid (see Figure 3 on page 17) (Oltvai and Barabasi, 2002). However, to climb to the top, we need to bridge the gap between individual modules and the way they are organised in scale-free modular structures. According to the CYGD (the Comprehensive Yeast Genome Database) (Guldener, et al., 2005), the number of proteins that participate in known protein complexes is 2750 whereas the sum of the sizes of these complexes (in terms of the number of participating proteins) is 8932. Thus, many of the protein modules may overlap, i.e. share proteins with each other, since proteins may participate in more than one module. This brings us to the final question that will be investigated in this thesis:

Question 6: Analysis of PINs reveals that some proteins appear in several modules. Can we find any patterns or common properties shared by these proteins? A positive answer to this question might be useful for revealing the role of modules in building higher-order structure(s) of the PIN organisation.

1.3 Contributions

The following contributions are made in this PhD thesis:

Contribution 1 We contribute a comparison of three semantic similarity measures that extends previous work by Lord, et al., (2003). In our extension we focus on different properties of the data sets that are relevant for the purpose of this work, such as the degree of clustering in the chosen data sets. This adds to our general understanding of the properties and usefulness of semantic similarity measures in large-scale PIN analysis.

Contribution 2 We contribute with a PIN clustering approach based on the combination of 1) topology information that regards shared

neighbourhood of the pair of proteins as a criterion for module membership and 2) the information content in GO annotation. The contribution consists of the procedure for merging of clusters based on the highest agreement between mentioned properties, which gives rise to modules that contain overlapping proteins. This reveals additional knowledge that is missed by methods that produce disjoint clusters based solely on topology information.

Contribution 3 We contribute novel protein-similarity measures that integrate semantic information and information based on PIN-topology. We describe several effective metrics that arise from this and demonstrate their use in analysing the properties of PINs.

Contribution 4 We develop a framework for identifying functional modules that utilises the proposed measures. Furthermore, we contribute with identifying the strengths and weaknesses of the combined methods for module identification, compared to the topology-based approaches that only consider triangle density measures.

Contribution 5 We find clear confirmation of the assumption that combining several biological aspects results in identification of more biologically plausible modules than using each aspect separately.

Contribution 6 We find evidence to suggest that considering the *k*-core sub-graph of the PIN in combination with the new measures (contribution 3) is more effective than applying those measures on the original PIN, hence adding to the growing evidence that *k*-cores form a useful concept in general analysis of biological networks.

Contribution 7 Using the measures and techniques developed in the thesis, particularly contributions 3-6, we find evidence supporting the following hypothesis: The proteins that take part in multiple modules (with a high number of occurrences) within the PIN, may be involved in the assembly and arrangement of cell structures to a greater extent than the proteins with lower numbers of occurrences across the generated module sets. This is proposed as a finding that reveals the role of modules in building higher-order structure(s) of the PIN organisation.

1.4 Generalisation of the module-identifying framework

There are many ways of measuring similarity between proteins. Our main proposal presented in this thesis considers protein similarity based on an integrated score that takes into consideration protein interaction data (as a topology source) and functional information based on semantic similarity. As pointed out previously, an ideal approach should take into consideration both temporal and spatial data, to be able to reflect the true dynamics of the cellular networks. It is therefore worthwhile to discuss how the methods presented here may be generalised to cope with several sources of information. Our module-identifying framework may be generalised by:

1) considering several sources of topological information
2) considering several sources of functional information

Topological information may refer to, for example, protein-protein interactions obtained from different experimental sources, such as Y2H and MS. However, this information may also be derived from different topological properties like clustering coefficient, edge betweenness, etc.

Besides semantic similarity values based on protein GO terms that we used in this work, there are many other sources of functional information that may be useful for predicting membership in protein complexes. One of the most prominent sources is gene expression data generated using various high-throughput platforms, such as microarrays. Expression profile correlation coefficients may, for example, be used to assign similarity scores to pairwise interactions. Other sources of functional information are essentiality, phylogenetic profiles, localisation, the MIPS functional catalogue, etc.

In this study, as in the majority of others, protein interactions are treated as binary, i.e. the edges in a network are either present or absent. Bearing in mind the fact that large-scale methods, although offering vast improvements in efficiency, still have much higher error rates than small-scale methods, a step towards generalisation of the proposed algorithms would be to treat protein interaction networks probabilistically. By treating the edges as binary (indicating presence/absence of interaction), we cannot distinguish edges supported by multiple evidence types, from edges supported by evidence of differing quality. There are several ways of assigning probabilities to individual pairs of proteins based on the amount and type of supporting evidence (Asthana, et al., 2004; Jansen, et al., 2002; Jansen, et al., 2003).

When dealing with several data sources that need to be combined in order to improve the prediction, a usual way of combining these consists of overlapping different interactomes. This approach, in turn, gives rise to the question whether it is more beneficial to consider the union of the disparate datasets or their intersection. As discussed in (Jansen, et al., 2002), one of the extremes that may be envisaged is that each one of the networks that are to be integrated have a low rate of false positives (FP) but a high rate of false negatives (FN). In this case, the union of the two sets of interactions would be advantageous. At the other extreme, when dealing with networks with high FP rates and low FN rates, the intersection between the different networks is preferable.

The problem of finding an optimal combination of unions and intersections among the different networks may be defined, as described in (Jansen, et al., 2002), as finding a trade-off between the highest possible coverage (TP/(TP+FN)) and the lowest possible error rate (FP/(TP+FP)). Determining the error rate is still an open question, as pointed out in (Jansen, et al., 2002).

An example of integrating different data sources that may be useful in generalising the proposed approaches is given in Figure 2 on page 13. The top part of the figure shows four possible data sources that may be useful for module identification. Two of them are topological sources, denoted as t_1 and t_2, and are usually treated as binary networks. The other two sources, denoted as f_1 and f_2 in Figure 2, may be used to assign functional weights to the edges. For example, when using gene expression as a possible source for weighting the edges, the probability of finding two proteins in a complex, given a certain correlation between their expression profiles, may be a possible way to assign weights (Jansen, et al., 2002). Gene ontology sub-graphs as a possible source of functional information is visualised in the third square in Figure 2, where semantic similarity between ontology terms may be used to reflect the functional similarity between the proteins, as assumed in this work. These functional weights may also be transformed into binary values, by setting different thresholds, where the level of the threshold determines the sensitivity and specificity of the experiment (see for example the transformation of semantic similarity values in Chapter 4). The bottom part of Figure 2 shows the hypothetical module sets generated with different combinations of data sets. The Venn diagram to the right in the figure shows binary subset profiles, where profile 1110 includes all data points that are present in data sets t_1, t_2, and f_1. $Mset_{1110}$, for example, denotes the set of modules derived from the combination of MS.

Y2H, and GO semantic similarity weights, where p_x denotes a protein x belonging the module.

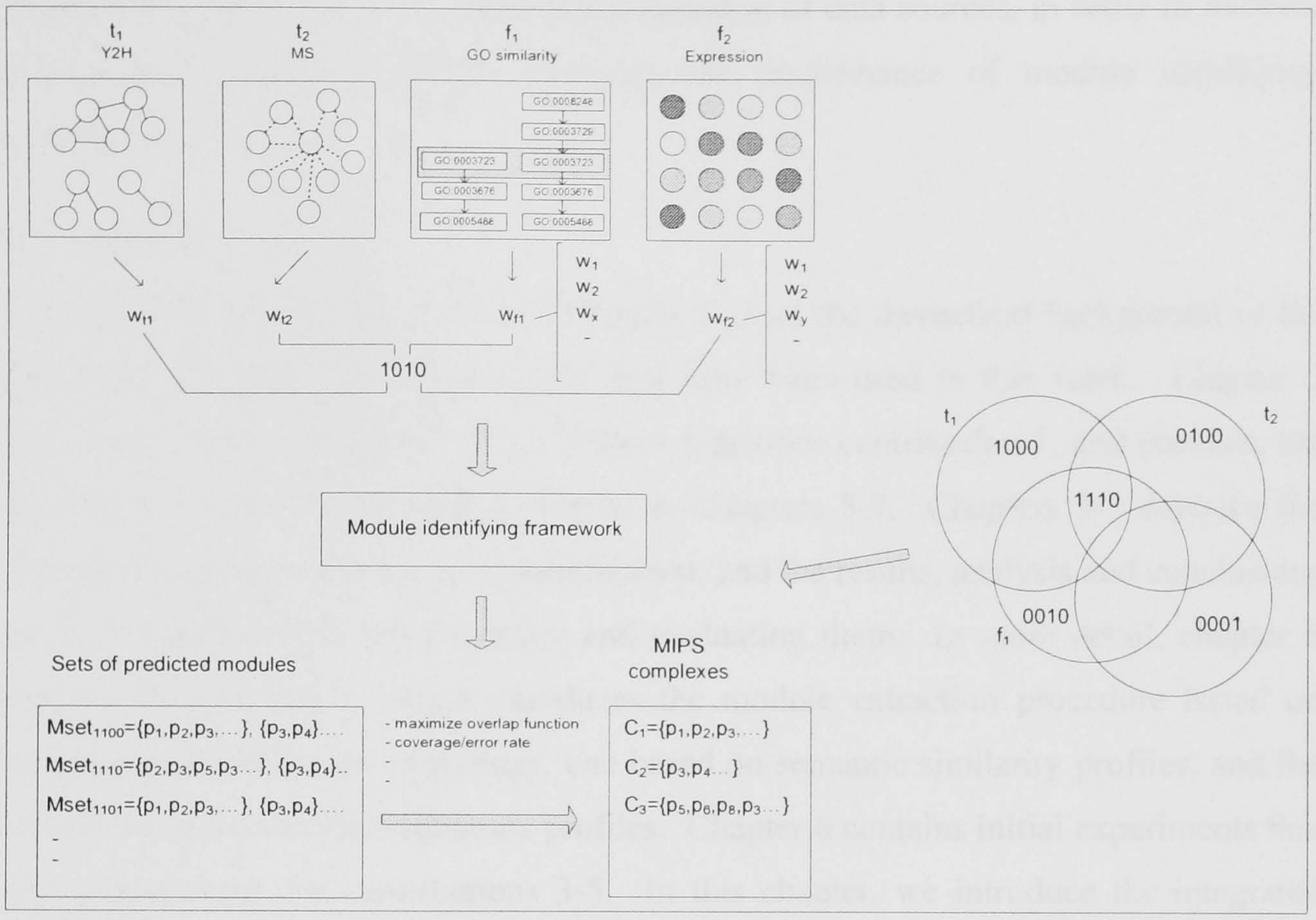


Figure 2: Hypothetical integration of four data sources for module identification

To be able to assess the prediction performance of various combinations of datasets, we need control datasets including gold standard positives (i.e. proteins that are connected) and gold standard negatives (i.e. proteins that are not connected). In this work, we have used the MIPS complex catalogue (Figure 2, bottom right) as gold standard to evaluate the predicted modules. Assuming this standard to determine whether two proteins belong to the same complex or not, different thresholds for predicting whether two proteins interact or not based on the datasets mentioned above could be tested, to optimise the performance of our module identifying framework. Besides error rate and coverage that we mentioned above, there is another measure that may be used to optimise the performance against the MIPS complex database, namely the overlap function, previously applied in (von Mering, et al., 2003). The overlap function is defined as TP/(TP+FP+FN). It is, however, important to bear in mind that the MIPS complex catalogue is not complete and that many FPs may be potential true hits worth investigating.

In (Asthana, et al., 2004) the probability that a given protein is in the same protein complex as a known core set of proteins may be defined as the probability that there

exists a path of direct and stable protein interactions between that protein and some member of the known complex. This probability could also be used, along with other mentioned sources, to find an optimal combination of data sources, in order to increase network reliability, and thereby improve the performance of module identifying algorithms.

1.5 Outline

This thesis is organised as follows. Chapter 2 gives the theoretical background of the thesis and introduces the data sources that have been used in this work. Chapter 3 describes related work in this area. Chapter 4 justifies contribution 1, and provides the foundation for the experiments presented in Chapters 5-7. Chapters 5-7 describe the proposed approaches for module identification, and the results, analysis and conclusions we have reached from implementing and evaluating them. In more detail, chapter 5 justifies contribution 2, which introduces the module extraction procedure based on integrating two different clusterings, one based on semantic similarity profiles, and the second based on mutual neighbours profiles. Chapter 6 contains initial experiments that give explanations for contributions 3-5. In this chapter, we introduce the integrated measures based on semantic information that bring in biological aspects into the topological measure. Furthermore, we investigate global network properties with these measures, present a framework that uses one of these measures for module identification, and finally compare the results from using different biological aspects in terms of the number of identified modules that match the known protein complexes. In Chapter 7, we describe the experiments on testing the effects of using the *k*-core aspect of the graph on module identification, which justifies contribution 6. Also in Chapter 7, additional experiments that underpin contribution 5 have been described. In addition, this chapter describes the comparison between two different types of measures that integrate semantic similarity with clustering coefficients. The purpose of Chapter 8 is to justify contribution 7 by exploring the proteins that participate most frequently in different modules (multi-modular proteins), and compare them to the proteins that are assigned to only one module (single-modular proteins). In Chapter 9, the modules obtained with our proposed approach are compared to other module sets generated by two other approaches, both based on topological properties. Advantages and disadvantages with each approach are also discussed in this chapter. Finally, the

conclusions drawn from this work and some ideas for future work may be found in Chapter 10.

Chapter 2

Background

2.1 Climbing life's complexity pyramid

Biological networks are often modular and compound, and involve connections between groups of genes and proteins as well as between individual elements. A simple complexity pyramid (see Figure 3 on page 17) suggested by Oltvai and Barabasi (2002), illustrates different levels of cellular organisation.

Living systems are organised at both logical and physical levels. The individual nucleotides are elementary building blocks of DNA and RNA molecules, which, in turn, are organised into higher level structures such as binding sites, regulatory elements, and genes. DNA is physically organised into larger structures such as chromatin and chromosomes. Groups of genes, proteins, RNAs, and metabolites (which are placed at the bottom level of the pyramid in Figure 3) may be organised into recurrent patterns, called pathways in metabolism, and motifs in genetic regulatory networks (level 2 in Figure 2). Regulatory motifs and metabolic pathways may in turn serve as building blocks of functional modules (level 3 in Figure 3). There is a growing body of evidence that the modules are then organised in a hierarchical manner (Barabasi and Oltvai, 2004; Oltvai and Barabasi, 2002; Ravasz and Barabasi, 2003; Ravasz, et al., 2002), defining the large-scale functional organisation of the cell (level 4 in Figure 3).

The way these various structures interact with each other determines the machinery of a cell. Cells and the extracellular matrix, which surrounds and supports cells, build up the tissues that in turn are organised into organs, and so forth.

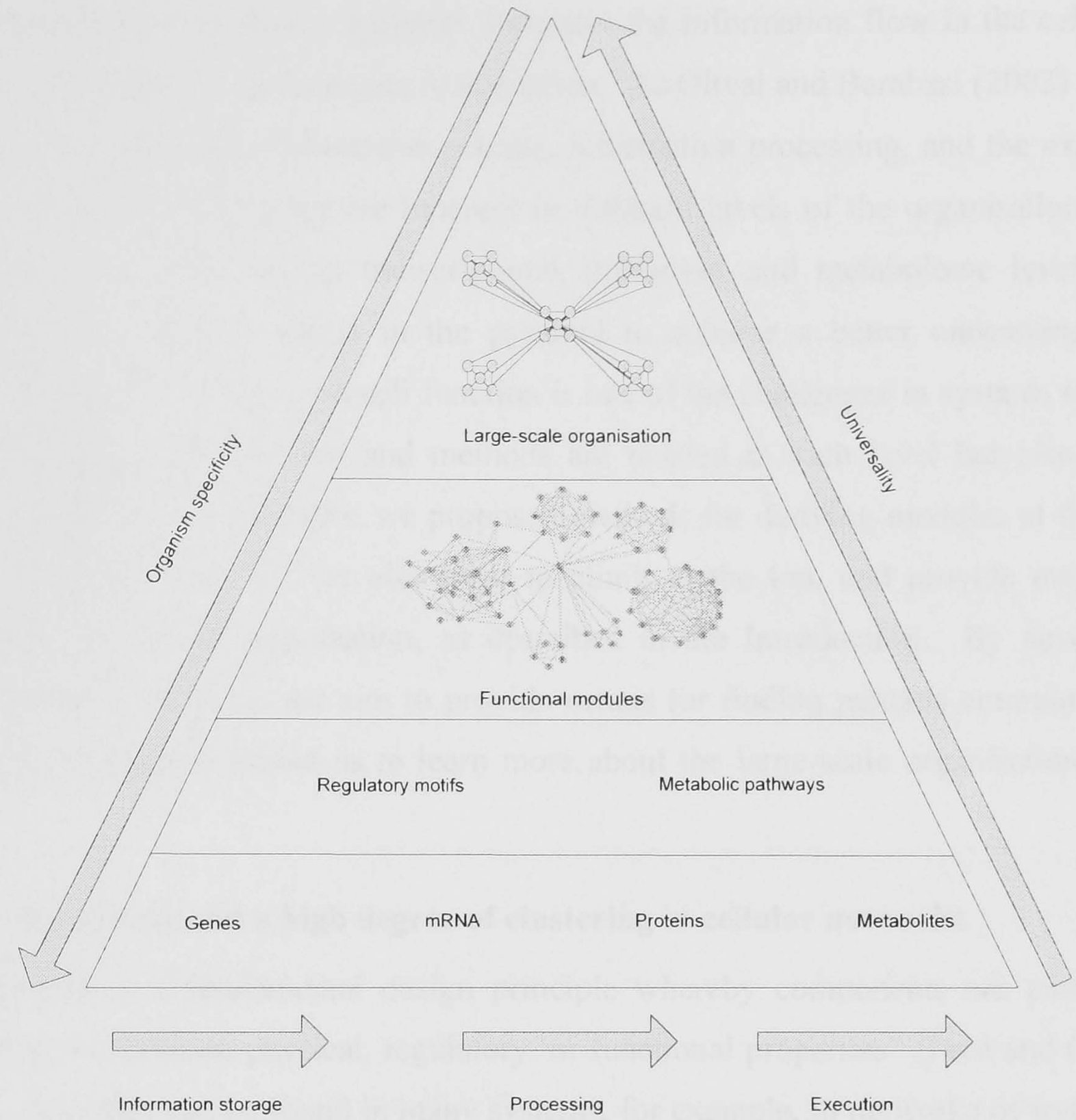


Figure 3: Life's complexity pyramid redrawn from (Oltvai and Barabasi, 2002)

Even though each organism has its unique components, biological networks are similar in structure to other complex networks, such as the World Wide Web (Broder, et al., 2000) and scientific co-authorship and citation networks (Newman, 2001; Redner, 1998). This similarity increases gradually from the particular (at the bottom of the pyramid) to the universal (at the top), as illustrated in Figure 3 above. While only a small proportion of metabolites (at the bottom level of the pyramid) are shared across different species, a higher degree of universality is found at level 2, i.e. metabolic pathways and motifs are frequently shared between species. An increasing level of universality is expected at the module level, and in the way modules are organised to form a scale-free architecture with inherent hierarchical relationships (Oltvai and Barabasi, 2002). Such large-scale architecture is also shared with some non-biological networks (see previous examples). This suggests that such non-random global organisational patterns are universal and apply to a wide range of complex networks (Oltvai and Barabasi, 2002).

Furthermore, the complexity pyramid illustrates the information flow in the cell: from information storage via processing to execution. As Oltvai and Barabasi (2002) state, it is widely accepted that information storage, information processing, and the execution of various cellular programs are inherent in different levels of the organisation of the cell, namely at the genome, transcriptome, proteome, and metabolome level. The integration of different layers in the pyramid to achieve a better understanding of system-level rules that govern cell function is one of the challenges in systems biology. Computational analysis tools and methods are needed at each level but also across different levels. In this thesis, we proposed methods for deriving modules at the third level in the pyramid, but we also want to climb to the top, and provide means for revealing large-scale organisation, as described in the Introduction. By developing modular representations, we aim to provide means for finding patterns common to the network, which will enable us to learn more about the large-scale organisation of the cell.

2.2 Modularity and a high degree of clustering in cellular networks

"Modularity is a fundamental design principle whereby components are partitioned according to common physical, regulatory, or functional properties" (Petti and Church, 2005). Modules can be found in many systems, for example, in networks of web pages describing related topics (Flake, et al., 2002), networks of friends in sociology (Newman, 2003), or scientific collaboration networks (Newman, 2001). A usual synonym for the term module in other scientific disciplines, like sociology for example, is *community* or *community structure*. In a study by Flake et al., (2002), the term web community is for example defined as "a collection of web pages such that each member page has more hyperlinks within the community than outside of the community". They state further that this definition may be adjusted to identify communities of varying sizes and levels of cohesiveness (clustering). An excellent review of the algorithmic methods for identifying communities of densely connected nodes in large networks may be found in (Newman, 2004).

Molecular biology is a highly modular science where functional modules are considered to be a critical level of biological organisation (see Chapter 1). The concept of a module in molecular biology was originally defined in (Hartwell, et al., 1999) as a discrete unit with a function that is separable from those of other modules. The separability may originate from, for example, cellular localization or specific protein

interactions. Furthermore, modularity involves groups of elements that work in a co-operative fashion to achieve some well-defined function.

In a general network representation, a module appears as a highly interconnected group of nodes (Barabasi and Oltvai, 2004). Modules can be interpreted as separated substructures of a network or pathway, e.g. a protein complex is a module of a protein interaction network. Protein complexes are well-defined examples of modularity since they consist of proteins that interact functionally and physically to form a robust unit, which, in turn, carries out some biological function (Yook, et al., 2004). Another example of modular organisation can be found in genetic regulatory networks where several transcription factor binding sites, organised into functional units, i.e. modules, play a crucial role in gene transcription (Klingenhoff, et al., 2002). Another property of modules is that their members are more strongly related to each other than to members of other modules, which is reflected in the network topology.

The modular nature of the cellular networks, including PINs, is reflected by a high degree of clustering, measured by the clustering coefficient. The clustering coefficient measures the local cohesiveness around a node, and it is defined, for any node i, as the fraction of neighbours of i that are connected to each other (Watts and Strogatz, 1998). Simply stated, the clustering coefficient c_i reflects the presence of 'triangles' which have a corner at i (see the triangles with dashed sides in Figure 4 below). The high degree of clustering is based on local sub-graphs with a high density of internal connections, while being less tightly connected to the rest of the network (Uhrig, 2006).

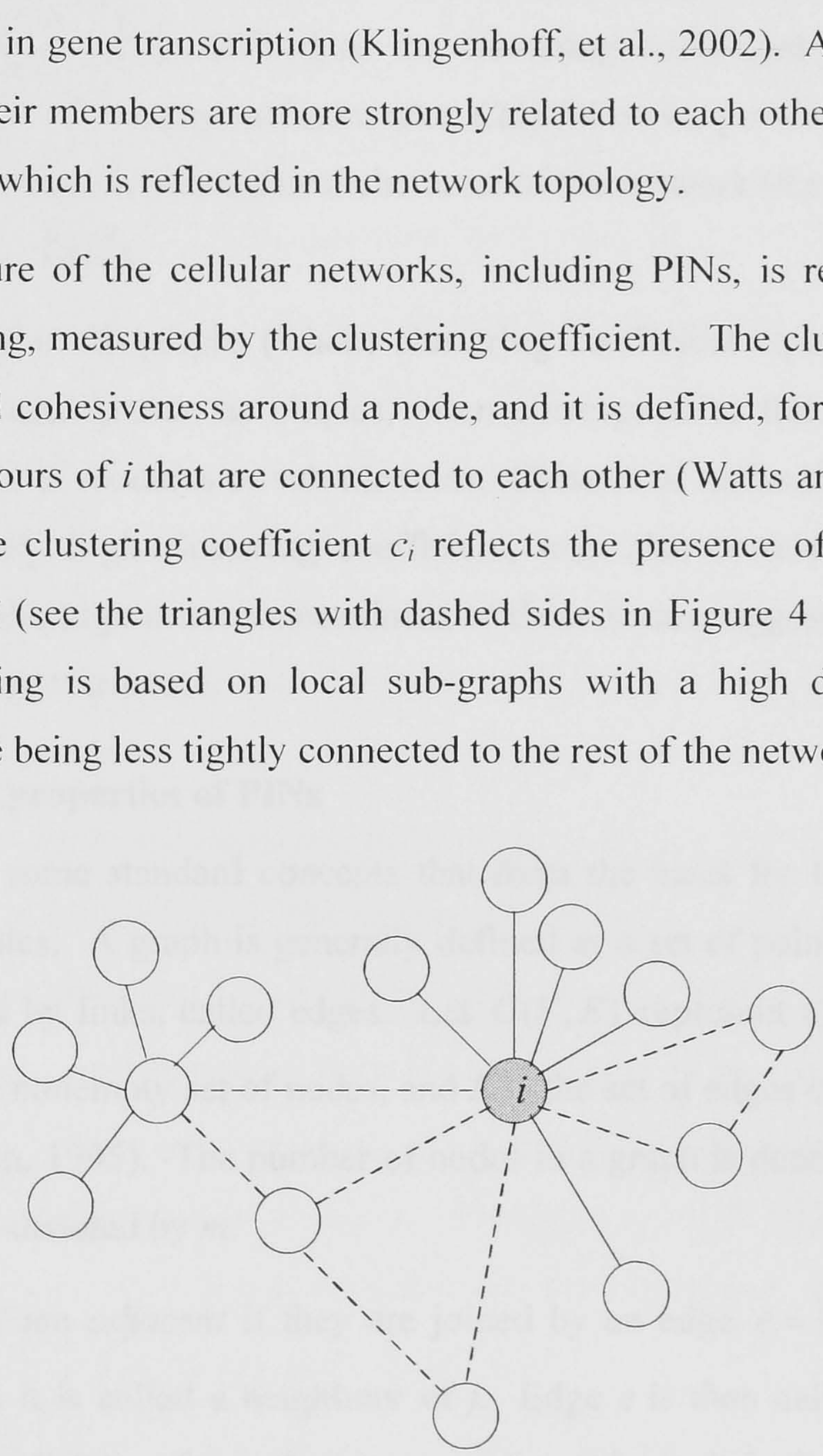


Figure 4: Example of a protein sub-graph with triangle-forming proteins

As pointed out by Barabasi and Oltvai (2004), each module may be reduced to a set of triangles, and a high density of such triangles is highly characteristic for PINs, pointing at the modular nature of such networks. By averaging the clustering coefficient over all nodes we can obtain a global measure of the cohesiveness of the network, where a high average clustering coefficient C indicates the presence of modularity. It has been confirmed in many studies that most real large-scale networks tend to contain dense clusters, in the sense that the average clustering coefficient of such networks is much greater than for random networks. In contrast, if modularity is absent in the network, the average clustering coefficient is comparable to that of a randomised network.

Further information about the architecture of the network may be obtained by inspecting the clustering function $C(k)$, defined as the clustering coefficient averaged over all nodes with degree k. For many real networks, $C(k)$ follows a power-law decaying as a function of k, reflecting the hierarchical character of the network (Ravasz and Barabasi, 2003; Ravasz, et al., 2002).

Maximally connected sub-graphs (whose clustering coefficient equals 1, which is the maximum value), also known as cliques, often correspond to functional complexes, which are examples of modules or sub-modules. Clusters of nodes that are not cliques, but have sufficiently high clustering coefficient, may also correspond to functional modules, because the high interconnectedness of these nodes suggests that they form a functionally important structure.

2.3 Topological properties of PINs

Here we describe some standard concepts that form the basis for the graph-theoretic definition of modules. A graph is generally defined as a set of points, called nodes or vertices, connected by links, called edges. Let $G(V, E)$ represent a simple undirected graph where V is a nonempty set of nodes, and E is the set of edges connecting a subset of the nodes (Rosen, 1995). The number of nodes in a graph is denoted by n while the number of edges is denoted by m.

Two nodes i and j are *adjacent* if they are joined by an edge $e = \{i, j\}$. If node i is adjacent to node j, it is called a neighbour of j. Edge e is then called *incident* to the nodes i and j. The *degree* of a node i in an undirected graph is the number of edges incident to i and it is denoted by k_i.

The neighbourhood $N(i)$ of a node i, consists of all neighbours of i, i.e. $N(i) = \{j \in V \mid \{i, j\} \in E\}$. By the closed neighbourhood $N[i]$ of node i, we mean $N(i) \cup \{i\}$. For the purpose of this work, we also define the set of edges connecting neighbours to node i as $K(i) = \{\{j,k\} \mid \{i,j\} \in E \wedge \{i,k\} \in E\}$. The number of edges that connect neighbours of node i to each other is denoted as n_i.

2.3.1 Node degree and distribution

Whereas node degree is one of the basic characteristics of individual nodes, *degree distribution* can be used to characterise the whole network. The degree distribution $P(k)$ is defined as the fraction of nodes that have degree k, and is obtained by counting the number of nodes that have $k = 1,2,3,...$ incident edges and dividing it by the total number of nodes N (Barabasi and Oltvai, 2004).

The degree distribution of many complex networks, such as the WWW, social networks, and cellular networks, follows a so called power law, i.e., $P(k) \sim k^{-\gamma}$, where γ is the degree exponent (Barabasi and Oltvai, 2004). This function indicates that the network does not have a characteristic node degree, like random graphs for example. Because of the absence of a characteristic scale, these networks are called *scale free* (see further 2.4.2). In contrast, *random* networks (see further 2.4.1) have a bell shaped degree distribution, which peaks at the average degree and decreases fast for both smaller and larger degrees.

Degree exponent γ indicates the importance of the hubs, i.e. highly connected nodes (see further 2.5.2) in the network. The smaller the value of γ, the more important the role of the hubs (Barabasi and Oltvai, 2004). For $\gamma > 3$, scale-free networks behave like random networks, and the role of hubs is not relevant. If the value of γ lies within the interval [2, 3], it indicates the presence of a hierarchy of hubs, where the largest hub is connected to a small fraction of all nodes. Finally, a network architecture known as *hub-and-spoke* emerges when $\gamma = 2$, meaning that the largest hub is connected with a large fraction of all nodes (Barabasi and Oltvai, 2004).

2.3.2 Clustering coefficient and distribution

The clustering coefficient measures the local cohesiveness around a node and it is defined, for any node i, as the fraction of connected neighbours of i (Watts and Strogatz. 1998):

$$c_i = \frac{2n_i}{k_i(k_i - 1)} \quad (1)$$

where n_i denotes the number of direct links between the k_i neighbours of node i. Simply stated, the clustering coefficient c_i reflects the presence of 'triangles' that go through node i (see the triangle with dashed sides in Figure 4 on page 19).

Furthermore, by averaging the clustering coefficient over all nodes we can obtain a global measure C of the cohesiveness of the graph, where a high average clustering coefficient indicates the presence of modularity, i.e. the tendency of the network to form clusters. It has been confirmed in many studies that most real large-scale networks have a tendency to cluster, in the sense that C is much greater than for random networks of equal size (which is approximately $\bar{k}/n$, where $\bar{k}$ is the average degree). The high interconnectedness of the nodes that belong to a certain sub-graph suggests that the sub-graph forms an important structure. Protein complexes constitute one example of such a structure in protein networks.

Besides the average clustering coefficient, another important network measure is clustering function $C(k)$, which is defined as the average clustering coefficient of all nodes with degree k (Barabasi and Oltvai, 2004). Values of this function show that $C(k)$ is independent of k in both scale-free and modular networks. However, for metabolic networks, it has been shown that $C(k)$ is well approximated by $C(k) \sim k^{-1}$, which according to Ravasz et al., (2002) provides evidence for a hierarchical structure of the network. The hierarchical topology brings together scale-free and modular topologies. As a consequence of this, it is further suggested in (Ravasz and Barabasi, 2003) that we should not think of modularity as the coexistence of relatively autonomous groups of nodes, as suggested earlier. Instead, this type of network is characterised by many small clusters that are tightly interconnected. These are combined in an iterative manner to form larger, but less interconnected groups (Ravasz and Barabasi, 2003).

2.3.3 Mutual clustering coefficient

A high degree of clustering in networks indicates that neighbours of a node are more likely to be neighbours of each other (i.e. to have edges between them) than would be expected in a random graph. For example, for an edge *ij* between nodes *i* and *j* (see Figure 5 below), a neighbour of *i* is more likely to have an edge to *j* if the edge is from a graph with a high degree of clustering (such as a small-world graph, see section 2.4.3) than if the edge is from a random graph. Hence, the mutual clustering coefficient measures the neighbourhood cohesiveness around individual edges (Goldberg and Roth, 2003) and not around nodes (Watts and Strogatz, 1998), like its predecessor described in section 2.3.2.

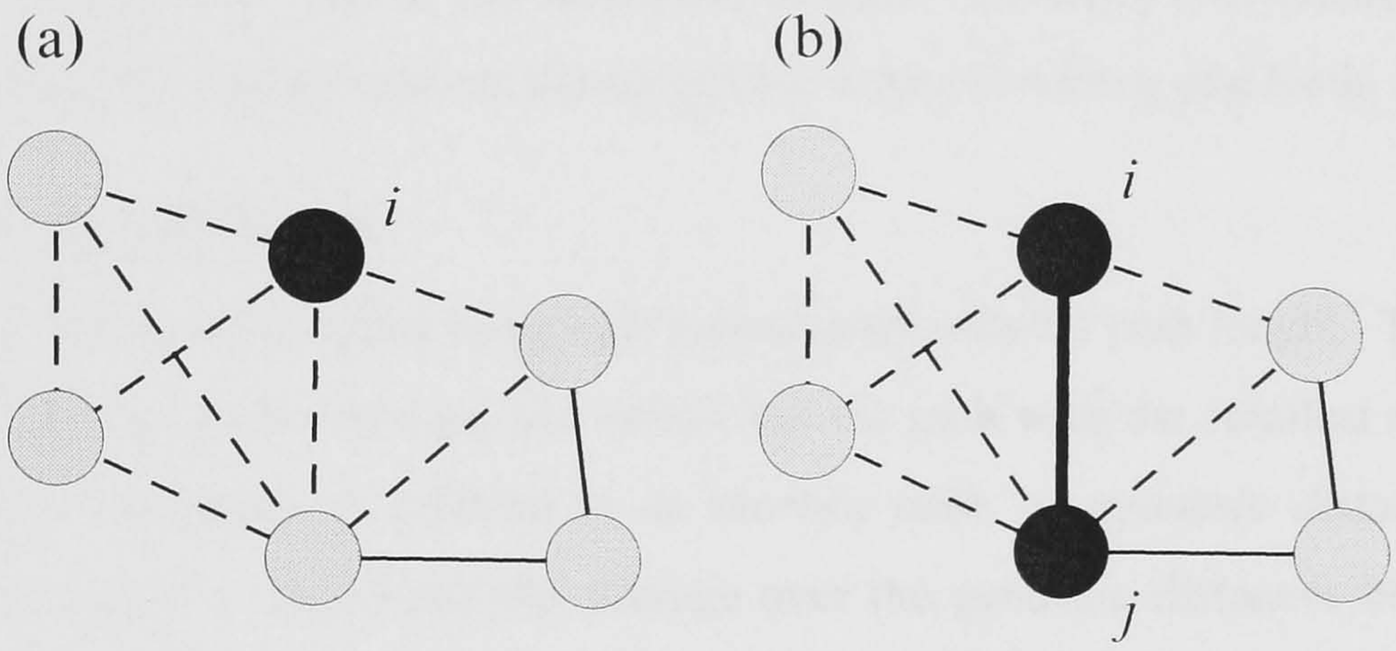


Figure 5: Examples of cohesive neighbourhoods. (a) Illustration of clustering coefficient. The neighbours of node *i* are more likely to be neighbours of each other (forming triangles with dashed lines) in a PIN than in a random network. (b) Mutual clustering coefficient. The two nodes connected by edge *ij* are more likely to share common neighbours in a PIN than in a random network

For two nodes, *i* and *j*, which are connected by an edge, the mutual clustering coefficient is defined as (Goldberg and Roth, 2003):

$$c_{ij} = \frac{|N(i) \cap N(j)|}{\min(|N(i)|, |N(j)|)} \quad (2)$$

where $N(x)$ denotes the neighbourhood of a node x. Mutual clustering coefficient is also called Meet/Min clustering coefficient. Goldberg and Roth (2003) also describe two alternative measures of mutual clustering coefficient, similar to the Meet/Min clustering coefficient. All three measures share a common numerator, but differ in their normalisation factors. For example, the Jaccard Index uses the union of the sets of edge

neighbours as normalisation factor, i.e. $|N(u) \cup N(v)|$ and the Geometric coefficient uses the product of the sets of edge neighbours.

Meet/Min clustering coefficient is chosen in this work for the purpose of module identification (see Chapter 5) since it is particularly suited for scale-free networks. If two interacting proteins share 15 neighbours, for example, we might want to put more weight on the mutual clustering coefficient of an edge if one of the proteins has only these 15 neighbours and the other has 125, than if each protein has 70 neighbours. Situations like this are expected in scale-free networks. By using the size of the smaller neighbourhood as normalisation factor, like in Meet/Min coefficient, the corroboration of such a pair of nodes will be weighted more heavily, than if the Jaccard coefficient was used, for example. All of the mentioned mutual clustering coefficients increase with the increasing overlap between the neighbourhoods (Goldberg and Roth, 2003).

2.3.4 Average path length

The distance between the nodes in a graph is measured with the path length. There may be many alternative paths between two nodes, but the path with the smallest number of links between two nodes is referred to as *shortest path*, or *geodesic distance*. The *average path length* ℓ represents the average over the geodesic distances between all pairs of nodes, and is defined as (Newman, 2003):

$$\ell = \frac{2}{n(n+1)} \sum_{i \geq j} g_{ij} \tag{3}$$

where g_{ij} is the geodesic distance from node i to node j.

The average path length indicates how efficiently information can be transmitted over the network. This network property gave the name to the particular class of networks, called small worlds (see section 2.4.3) in analogy with the concept of the small-world phenomenon (originally observed in social systems), that is characterised by high clustering and small average path length.

2.4 Network models

We describe here some common and widely used network models, such as random, scale-free, and small-world networks (see sections 2.4.1-2.4.3). One recently emerged network model, named hierarchical network is described in section 2.4.4. We also point

out the basic properties of these networks. More thorough surveys of large network models may be found in (Albert and Barabási, 2002; Newman, 2003).

2.4.1 Random networks

The random graph model relies on the basic principle that the probability p that there is an edge between any pair of nodes is distributed uniformly at random. In their first paper on random graphs, Erdős and Rényi defined a random graph as a graph with n nodes connected by m edges, which are chosen randomly from $n(n-1)/2$ possible edges (Erdős and Rényi, 1959). The node degrees for such networks follow a Poisson distribution indicating that the majority of nodes have approximately the same degree, which is close to the average degree of the network (Barabasi and Oltvai, 2004).

Even though random graphs have been successfully used to model some aspects of real networks (see for example the randomly constructed genetic networks in (Kauffman, 1969)), some of their properties differ from the properties of real networks and they cannot be used to plausibly approximate those. As pointed out in section 2.3.1, many real networks have a power-law degree distribution (Albert, 2005; Albert, et al., 2000; Barabasi and Oltvai, 2004), with the consequence that the probability of finding nodes that are highly connected (also known as hubs, see further section 2.5.2) is significantly higher than in random graphs (for more details about degree distribution, see 2.3.1). Hence, the absence of any dominant hubs in random networks is one of the features that differ from real world networks (Strogatz, 2001).

The second difference is that random graphs, such as those defined by Erdős-Rényi are poorly clustered, i.e. have much lower average clustering coefficients compared to the real world networks. For example, in the neural network of *C. elegans*, the clustering coefficient is $C = 0.28$ while the corresponding value for a random network with the same number of nodes and average number of edges per node is only 0.05 (Watts and Strogatz, 1998).

Random graphs exhibit small average path length, which is proportional to $\log N$ and indicates the presence of small-world property in such graphs.

2.4.2 Scale-free networks

In many real networks, some nodes have a much higher degree than others (Strogatz, 2001). Some of the example networks whose degree distribution decays according to a

power-law, i.e $P(k) \sim k^{-\gamma}$, are metabolic networks (Jeong, et al., 2000) and the World Wide Web (Broder, et al., 2000). This heavy-tailed degree distribution is characteristic for the scale-free network model that was proposed in (Barabasi and Albert, 1999). This distribution emerges from a stochastic growth model according to the "rich gets richer" principle, meaning that new nodes are added continuously and preferentially attach to existing nodes with a probability proportional to the degree of the target node (Strogatz, 2001). This means that the nodes with high degrees become even more connected, which results in the presence of hubs. The relevance of hubs in relation to degree exponent is described in section 2.3.1.

In a study by Cohen and Havlin (2003), the average path lengths for scale-free networks with varying degree exponent γ have been investigated. The results show that scale-free networks with $2 < \gamma < 3$ are ultra small (see 2.3.1 for a more detailed discussion on the relevance of hubs in such networks) with average path length behaving as $\ell \sim \log \log n$. In scale-free networks with $\gamma > 3$, ℓ approximates $\log n$, meaning that the scale-free networks behave like random networks in this aspect (Cohen and Havlin, 2003).

One of the theories that try to explain the mechanism that underlies the scale-free architecture of cellular networks is concerned with gene duplication, as described in Barabasi and Oltvai (2004). Genes that undergo duplication produce identical proteins that connect to the same protein partners. As a result of this, each protein that is linked to the duplicated protein gets an extra link. Hubs, which are highly connected, have a higher probability of being connected to the duplicated proteins, and are more likely to gain additional connections if the proteins to be duplicated are randomly selected (Barabasi and Oltvai, 2004).

The question that has been discussed by Strogatz (2001) in the context of the importance of scale-free architectures is: "Could there be a functional advantage to scale-free architectures?". It has been shown that scale-free networks are robust against accidental failures due to the presence of a few high-degree hubs that dominate the topology of the networks. Instead, the random failures that affect mainly small-degree nodes will not cause any severe damage to the network (Albert, et al., 2000). On the other hand, the dominance of hubs may also cause *attack vulnerability* (Albert, et al., 2000), meaning that the knock out of a few hubs partitions the network into small

isolated clusters. The design of therapeutic drugs and the evolution of metabolic networks are examples of possible implications of this property that have been discussed in (Jeong. et al.. 2000).

2.4.3 Small-world networks

Many real networks lie somewhere between the extremes of order and randomness (Strogatz, 2001). Such networks, found in biological, social. and many other systems, often exhibit a *small-world* topology. In small-world networks, any pair of nodes can be connected with a path of a few links. The overall navigability of the networks is measured by the average path length, which is defined in section 2.3.4. Besides short average path length, small-world networks are also characterised by unusually large clustering coefficients, independently of network size (Watts and Strogatz, 1998).

Although random networks also have short average path lengths ($l \sim \log n$). they cannot be classified as small-world networks because their average clustering coefficient is much smaller than the corresponding value of small-world networks.

Besides the pioneering work by Watts and Strogatz (1998) on the small-world property that is generic for many real networks, several additional empirical examples of this type of architecture have been explored (Amaral. et al.. 2000; Barabasi and Albert, 1999; Jeong. et al.. 2000).

2.4.4 Hierarchical networks

In a pioneering study by Ravasz et al., (2002), a new class of networks has been proposed. In their attempt to bring together modularity, a high degree of clustering and a scale-free architecture, they proposed a model that captures all of these features, namely the *hierarchical network* model. For this purpose, an assumption needs to be made that modules are combined with each other in a hierarchical way, thereby generating a hierarchical network (Ravasz and Barabasi, 2003; Ravasz, et al., 2002).

Such a network is characterised by a power-law degree distribution with a degree exponent that varies between 2 and 3 and a large, size-independent average clustering coefficient. The most important signature of hierarchical modularity is the scaling of the clustering coefficient, which approximates $C(k) \sim k^{-1}$ and may be seen as a straight line with a slope of –1 on a log–log plot. A hierarchical architecture implies that sparsely connected nodes are parts of more cohesive modules, which in turn are parts of

even more cohesive modules. There are often a few hubs that maintain communication between the different highly clustered parts of the network.

2.5 Biological interpretation of the topological properties

As pointed out earlier, the topological features of biological networks are shared to a great extent by other complex networks. While this universality allows us to apply graph theoretical measures to learn more about biological networks, it is important to not dismiss the interpretation of the topological properties in the light of functional and evolutionary information.

2.5.1 Modularity and modules

"Modularity might very well represent a general attribute of living matter, with *de novo* invention being rare and reuse the norm" (Gavin, et al., 2006). Cellular networks are, unlike random networks, thought to be modular, i.e. composed of functionally related components that often interact with each other (Hartwell, et al., 1999).

The exact meaning of modularity in biological networks depends on the network under consideration. For example, modules in protein networks are often seen as static molecular complexes (such as the ribosome) or as dynamic signalling pathways (such as the MAPK cascade). There are also examples of large modular molecule complexes that are in turn organised in modules. One of such complexes is yeast Mediator, which transmits regulatory signals from DNA-binding transcription factors to RNA polymerase II (Guglielmi, et al., 2004). The Mediator complex is thought to be composed of 24 subunits organised in four modules, named the head, middle, tail and Cdk8 modules. In gene regulatory networks, modules are often seen as sets of genes controlled by the same set of transcription factors under certain conditions (Segal, et al., 2003).

Modules should not be seen as isolated components, since it has been shown that some crosstalk and overlap exists between them (Han, et al., 2004; Schwikowski, et al., 2000). Instead, modules should be considered as components that have dense intra-connectivity but sparse inter-connectivity. In a study analysing protein interaction networks in the yeast *Saccharomyces cerevisiae*, Schwikowski et al., (2000) reported global patterns of interactions of proteins within functional classes or subcellular compartments, as well as many possible cross-connections. It is further pointed out by Qi and Ge (2006) that the existence of the links between modules emphasises the

coordination of the cellular processes. For example. Petti and Church (2005) investigated possible transcriptional coordination between glycolysis and lipid metabolism modules.

A growing body of work supports the idea that such modules underlie much of cellular functioning (Gavin, et al., 2006; Han, et al., 2004; Pereira-Leal, et al., 2004; Qi and Ge, 2006; Rives and Galitski, 2003), and that functional modules are the most relevant organisational units of a cell from the perspective of systems biology (Hartwell, et al., 1999).

2.5.2 Hubs

In scale-free networks, such as protein interaction networks, the probability of a node being highly connected is higher than in a random graph. Highly connected nodes, which often correspond to hubs, govern network properties to a great extent (Albert, 2005). One of the characteristics of scale-free networks is their heterogeneity, which is partly attributed to hubs. Because of this property, these networks are robust, meaning that they have a high tolerance to random perturbations, but also a high sensitivity to targeted attacks on highly connected nodes (Albert, et al., 2000). The failure of hub nodes causes a breakdown of the network into isolated clusters, while failure of random nodes mostly affects small-degree nodes, since they are the most abundant, and does not cause any major loss of connectivity (Albert and Barabási, 2002).

One intensively studied example of a hub protein is the p53 tumour suppressor (Vogelstein, et al., 2000). This hub has a role in cellular apoptosis through the regulation of several target genes. In 50% of human tumours, the p53 protein is inactivated directly as a result of mutation in the gene that produces it, while in the remaining tumours the role of p53 is inactivated indirectly through binding to viral proteins or as a result of alternation in genes whose products interact with p53 (Vogelstein, et al., 2000). This exemplifies the attack vulnerability caused by the reliance on hubs in protein networks.

It is further hypothesised that the importance of hubs may be explained by their evolutionary conservation. It has been shown by comparing putatively orthologous sequences between *S. Cerevisiae* and *C. elegans* that the connectivity of well-conserved orthologs is negatively correlated with their evolutionary rate (Fraser, et al., 2002). Hence, hubs are shown to have smaller evolutionary distances to their orthologs, and

such proteins evolve more slowly because "a greater proportion of the protein is directly involved in its function" (Fraser, et al., 2002). In the same study, it is confirmed that highly interacting proteins are more likely to be required for the viability of the network.

2.5.3 Motifs and cliques

In most real networks we can find different types of sub-graphs, such as squares, triangles, etc. Some types are overrepresented in comparison to the expected representation of the same type of sub-graph in a randomly generated network of the same size. Such graphs are suggested to form motifs which are topologically distinct interaction patterns in each real network (Milo, et al., 2002). Milo et al., (2002) define network motifs as "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomised networks". One of the most significant motifs in both *E. coli* and yeast is the feedforward loop (FFL). In FFL, transcription factor *X* regulates transcription factor *Y*, which in turn result in their joint regulation of gene *Z*. Other types of network motifs found in *E. coli* are single input motifs (SIM), and multiple input motifs (MIM).

Wuchty et al., (2003) showed that motifs are evolutionarily conserved. Another interesting result that this study indicates is that the different functions are not only associated with characteristic topological motifs, but that they also conserve these motifs at different rates during evolution.

Motifs in protein interaction networks may represent different types of interactions. Small fully connected motifs, also known as cliques, are abundant in such networks and have a tendency to form functional complexes (Wuchty, et al., 2003). In transcription regulation networks, transcription factor motifs are found to be abundant, as shown in the examples of *E. coli* and *S. Cerevisiae* where the motifs are results of the convergent evolution of the transcription regulatory networks of different species (Conant and Wagner, 2003). As further suggested by Conant and Wagner (2003), this is an indicator of optimal circuit design.

Network motifs provide, besides robust property of biological networks, an important tool for understanding the modularity and the large-scale structure of networks (Mangan, et al., 2003).

2.5.4 *K*-cores

As stated in section 2.5.2, a set of highly connected proteins play an important role for the network's integrity by connecting the inherent modules of the network. Such cohesive sub-graphs that correspond to modules may be found by applying a *k*-core decomposition of the network. Decomposition of the network into the core layer structure may also help finding topologically important proteins (Wuchty and Almaas, 2005).

The *k*-core of a graph is defined as the maximum sub-graph where every node has at least *k* links (see Figure 6 below). This sub-graph can be generated by recursively deleting the nodes with degrees lower than *k*, and their incident edges, until all nodes in the remaining graph have at least degree *k* (Bagatelj and Zaversnik, 2002). The *core number* of node *i* is defined in (Bagatelj and Zaversnik, 2002) as the highest order of a core that contains this node.

An algorithm for core decomposition of graphs is described in (Bagatelj and Zaversnik, 2002). The input is a list of all nodes and their neighbours (for example Node: R; Neighbours: H, M, N, P, see Figure 6), and the output is a table with core numbers for each node. The outline of the algorithm is given in Text box 1 on page 32. In Figure 6 below, we can see that cores are nested. Cores may also be disconnected sub-graphs (not shown here). The dark gray area comprises the nodes that build the sub-graph with the highest *k*, which here equals 3.

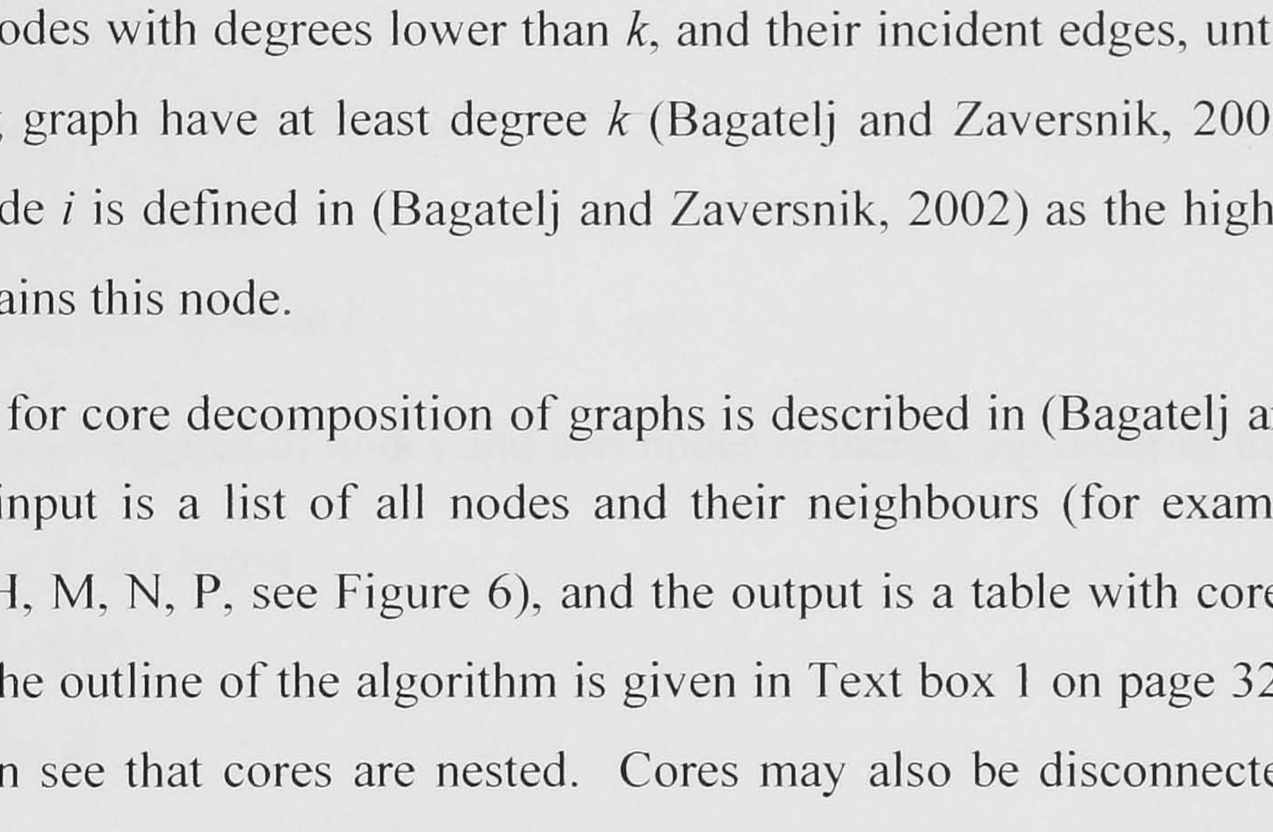


Figure 6: Decomposition of the graph into three core layers: 1, 2 and 3-core

This decomposition makes it possible to discern the most densely connected sub-graphs for each node. For example, node B has four neighbours: A, C, D, and E. Together with three of its neighbours: C, D, and E (which all belong to the 3-core), it forms a fully connected graph, while one of its neighbours (A) is isolated and belongs to the 1-core, and will therefore be discarded from the original sub-graph. Topological structures similar to the one formed by B, C, D and E are likely to correspond to protein complexes.

The outline of the algorithm for core decomposition (Bagatelj and Zaversnik, 2002)

▷ Input: $G = (V, E)$

▷ Output: *core* table

▷ $N(i)$: closed neighbourhood of i

▷ $core(i)$: *core* number

▷ k_i : connectivity of node i

Step 1: Compute degrees of nodes and sort nodes in increasing order of their degrees

For each $i \in V$ do begin

Identify $N(i)$;

Compute k_i;

end

Sort the elements of set V in the order of increasing k_i;

Step 2: Delete nodes with degree less than k

For each $i \in V$ do begin

$core(i) := k_i$;

For each $j \in N(i)$ do

If $k_j > k_i$ then begin

$k_j > k_j - 1$;

Sort the elements of set V in the order of increasing connectivity

end

end

Text box 1: The outline of the core decomposition algorithm

Methods based on *k*-cores have been used for the detection of molecular complexes in protein interaction networks (Bader and Hogue. 2003: Tong, et al., 2002). and determining the essentiality of proteins (Wuchty and Almaas, 2005). It has been shown that the probability of nodes both being essential and evolutionarily conserved increases towards the innermost cores (Wuchty and Almaas, 2005). It is further hypothesised in the same study that the proteins that belong to the innermost cores, also called globally central cores, serve as the evolutionary backbone of the proteome.

2.6 Domain knowledge

In this section, we aim to present the main sources of annotation that describes and structures the knowledge (using ontologies) about proteins, such as their membership in functional categories, involvement in molecular complexes, biological processes etc. We also describe the measures of semantic similarity that are used to calculate similarities between ontology terms in this work (see section 2.6.1).

2.6.1 Semantic similarity measures and Gene Ontology

One of the knowledge sources that are used in this work is Gene Ontology (2001). Gene Ontology (GO) offers a vocabulary of molecular biology for describing gene products in any organism. Individual terms are organised as direct acyclic graph (DAG), where the nodes denote the terms and the edges denote the relationships. Terms may be related to each other by two types of relationships, namely the "is-a" and "part-of" relationships. The ontology is sub-divided into three aspects: molecular function, biological process and cellular component. A tree-like structure of the GO terms for two gene products, BRR1 and SMX2 is illustrated in Figure 7 on page 34. We use the *Saccharomyces* Genome Database, SGD (http://genome-www.stanford.edu/Saccharomyces/) which contains GO annotations from all three sub-ontologies (Dwight, et al., 2002). In the example in Figure 7, each term is assigned a GO accession number. As an example, the term "RNA binding" (GO:0003723) is the most specific term assigned to BRR1, and it is a child term of the "nucleic acid binding" term (GO:0003676), which in turn has the term "binding" as an ancestor.

To calculate the semantic similarity between gene products, the probability of each term assigned to the gene product is first derived. The probabilities reflect how many times the term or any of its descendants occurs in the annotated database, in this case SGD. The probabilities are shown in each box in Figure 7 on page 34. The number in

parentheses denotes how many times the term or any of its descendants occurs in SGD. The procedure of calculating GO term probability, proposed by Lord et al., (2003), is described as follows. For each gene product, the probability is calculated by counting the number of times each term or its descendants occur in the annotations in SGD, divided by the total number of GO term annotations in SGD. The probabilities increase as we move towards the root, which is defined as "molecular function" (GO:0003674) and has probability 1. Given these probabilities, there are several ways to calculate semantic similarity (Jiang and Conrath, 1998; Lin, 1998; Resnik, 1999).

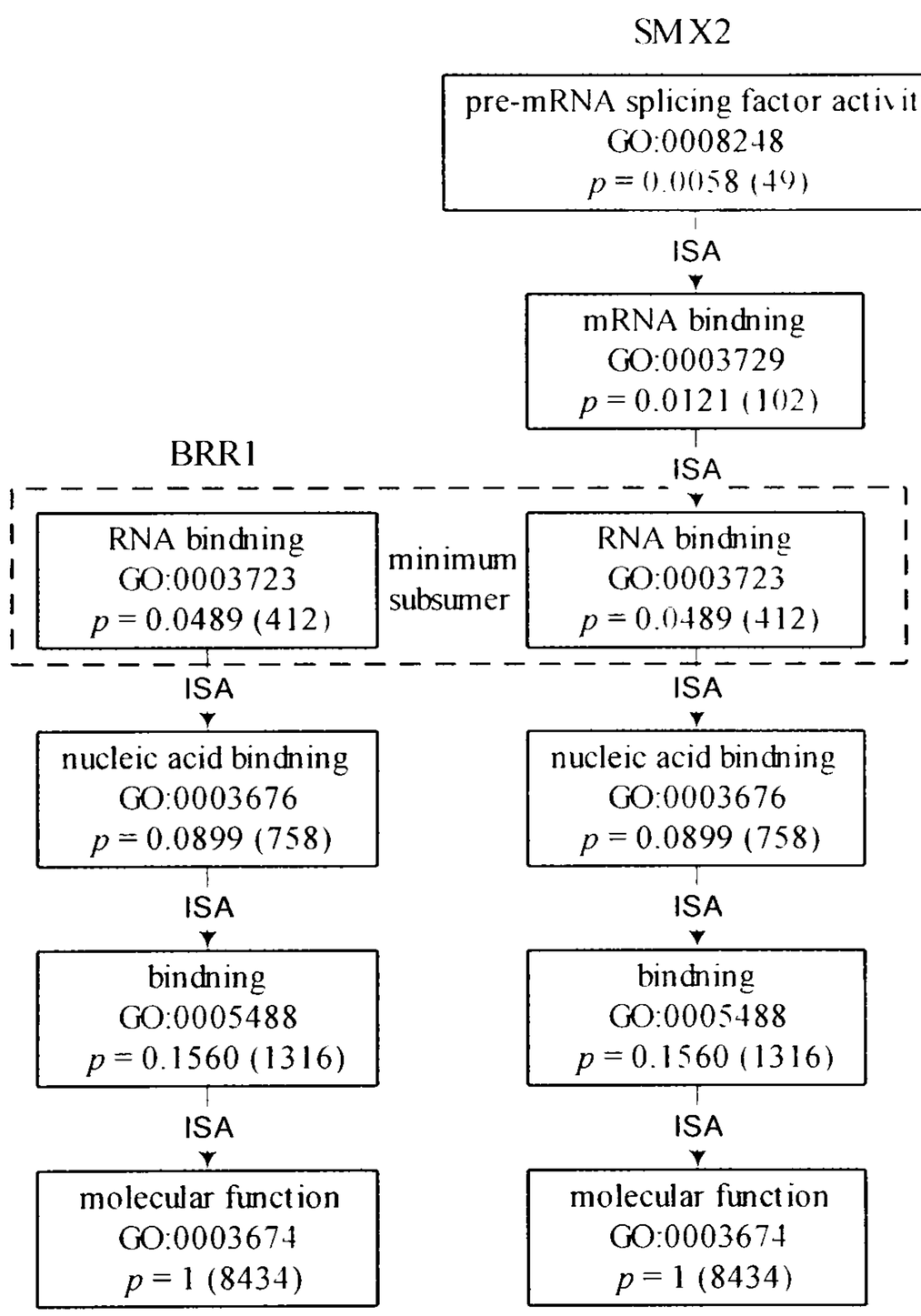


Figure 7: GO annotation sub-graphs describing GO molecular function terms for two example proteins

In order to calculate the similarity between two proteins i and j, we need to calculate the similarity between the terms belonging to the term sets T_i and T_j that are used to annotate these proteins. Given the ontology terms $t_k \in T_i$ and $t_l \in T_j$, the semantic similarity measure proposed by Lin (1998) is defined as:

$$sim(t_k, t_l) = \frac{2 \ln p_{ms}(t_k, t_l)}{\ln p(t_k) + \ln p(t_l)} \quad (4)$$

where $p(t_x)$ is the probability of term t_x and $p_{ms}(t_k, t_l)$ is the probability of the minimum subsumer of t_k and t_l, which is defined as the lowest probability found among the parent terms shared by t_k and t_l (Lord, et al., 2003). In Figure 7 on page 34, the minimum subsumer for BRR1 and SMX2 is "RNA binding" (GO:0003723).

As GO allows multiple parents for each term, two terms can share parents by multiple paths. Like Lord et al., (2003), we used the average term-term similarity since each protein can be annotated by several terms, and because we are here interested in the overall similarity between the pair of proteins rather than between pairs of individual ontology terms. The average term-term similarity can vary between 1 (for identical terms) and 0 (no similarity). Given two proteins, i and j, with T_i and T_j containing m and n terms, respectively, the protein-protein similarity is defined as the average inter-set similarity between terms from T_i and T_j :

$$ss_{ij} = \frac{1}{m \times n} \sum_{t_k \in T_i, t_l \in T_j} sim(t_k, t_l) \quad (5)$$

where $sim(t_k, t_l)$ is calculated using Equation 4 above.

The semantic measure proposed by Resnik (1999) only uses the information content of the shared parents. As p_{ms} varies between zero and one, this measure generates values between infinity (for similar terms) and zero. In practice, for terms that are present in the corpus (a large and structured set of texts), the maximum value is defined by $-\ln(1/t) = \ln(t)$, where t stands for the number of occurrences of any term in the corpus (Lord, et al., 2003). The semantic similarity measure proposed by Resnik (1999) is defined as:

$$sim(t_k, t_l) = -\ln p_{ms}(t_k, t_l) \quad (6)$$

The measure proposed by Jiang and Conrath (1998) is actually a semantic distance measure rather than a semantic similarity measure. It is, like the measure by Lin, based on both the information content of the parent terms and of the terms that are being compared, but uses the terms in a different order. In theory this measure can, as the

measure by Resnik, give arbitrarily large values, but in practice the maximum value is $2\ln(t)$ where t denotes the number of occurrences of any term in the corpus (Lord, et al., 2003).

Semantic distance between two terms according to Jiang and Conrath (1998) is defined as:

$$dis(t_k, t_l) = -2\ln p_{ms}(t_k, t_l) - (\ln p(t_k) + \ln p(t_l)) \quad (7)$$

2.6.2 Semantic similarity and functional homogeneity

Besides including semantic similarity into quantitative network measures and thereby considering the functional homogeneity of potential modules, we also provide a visual representation of semantic similarity by using these values as input to an existing graph visualisation tool.

For the purpose of visual analysis, we have created protein interaction graphs enriched with the semantic similarity weights, which are reflected by the width of the edge. The semantic weights that have been incorporated into the graph representation may result in increased confidence in functional relevance of the interactions (see Figure 8 on page 37). To generate a two-dimensional graph layout for protein networks, we use the GraphViz software (http://www.research.att.com/sw/tools/graphviz) developed at AT&T research labs. GraphViz starts from a textual description of the graph structure and is capable of generating graphs in a variety of output formats. An example of a protein interaction sub-graph that is enriched with semantic similarities is shown in Figure 8. In this figure, nodes represent interacting elements (proteins) and links denote interactions. In this example, we use semantic similarity based on molecular function as input to GraphViz. Nodes are coloured green if they are annotated with at least one term from GO molecular function sub-ontology, otherwise red. The width of each edge is proportional to the semantic similarity between the nodes it connects. The width can vary in the interval $[0,10]$, where 10 stands for the maximum similarity. Zero-width lines are replaced by dashed lines. In the example in Figure 8, the semantic similarity between PPH21 and PPH22 equals 1, which gives the maximal width of the edge (10), and the semantic similarity between TPD3 and YOR1 is 0.298, which gives an edge with the width 2.98. Node ZDS2 has unknown function and it does not have any functional similarity with the rest of the cluster, which results in dashed lines.

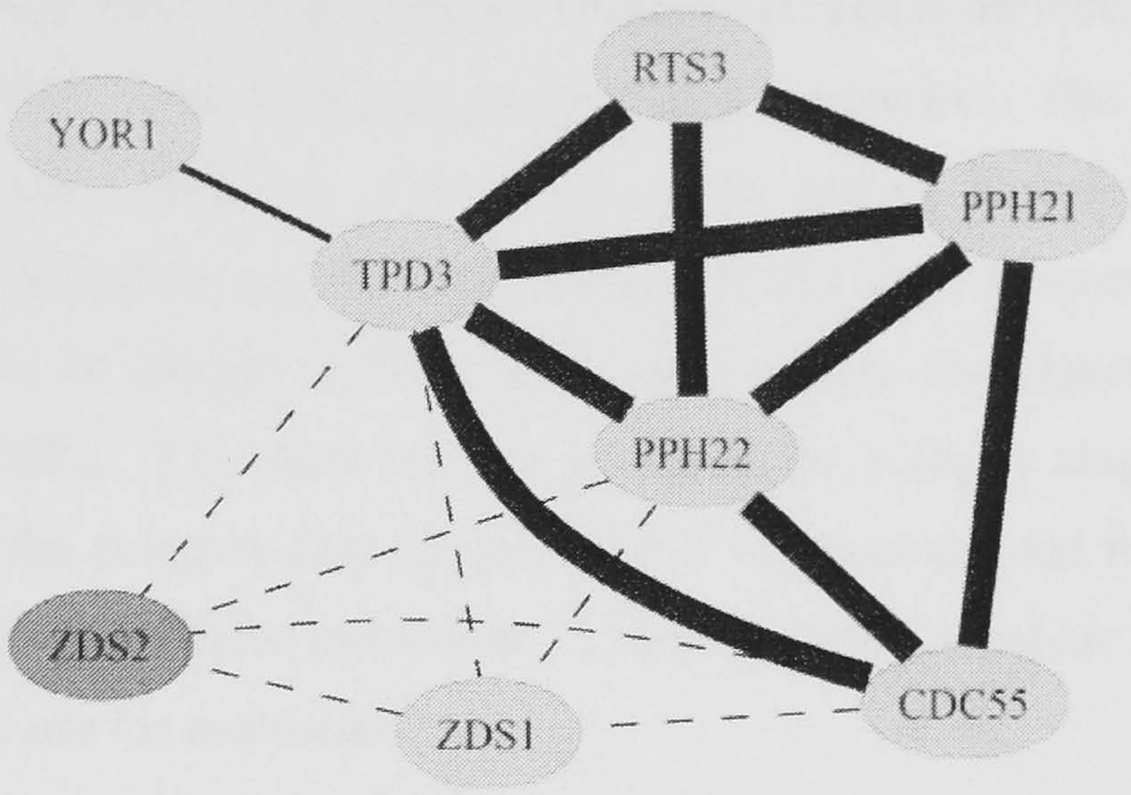


Figure 8: Graph representation for the protein TPD3 and its neighbours

By visual analysis of the sub-graph in Figure 8 above, we can discern a "clique" containing the proteins RTS3, TPD3, PPH21, PPH22, and CDC55, which are all connected with maximum similarity (except that the nodes RTS3 and CDC55 lack a direct connection between each other). This sub-graph represents the protein phosphatase 2A complex, which is a family of well established serine/threonine phosphatase complexes annotated with the GO-term "protein phosphatase type 2A activity". The remaining three proteins are not part of this activity and they have very low or no semantic similarity with the other proteins in the clique. Hence, by incorporating semantic similarity, we can distinguish a cluster containing a complex of strongly functionally related proteins and remaining interacting proteins, which is not obvious from the initial sub-graph of interactions.

2.6.3 MIPS

The Munich Information Center for Protein Sequences (MIPS) provides high quality curated genome-related information, such as protein-protein interactions, protein complexes, protein functional categories, etc., spanning over several organisms.

The MIPS functional catalogue database consists of different fields, such as functional catalogue (FunCat) number, EC number, GO number, keywords etc. FunCat is an annotation scheme that provides functional descriptions of proteins (Ruepp, et al., 2004). There are in total 28 main functional categories that are hierarchically structured.

The MIPS Comprehensive Yeast Genome Database (CYGD) provides information on the molecular structure and functional network of *S. cerevisiae*. The information that we use for evaluation purposes in this work is the protein complex catalogue that contains a manually curated set of protein complexes that serve as an example of a type of module. There is another data set containing protein complexes obtained from (Gavin, et al., 2002). This data set was produced by using a single experimental method. whereas the complex data set from MIPS has been derived from experiments from many labs using different techniques. Therefore. MIPS database is more realistic and appropriate to use for evaluation.

2.7 Protein interaction data

Since most of the protein-protein interactions used in this work come from yeast two-hybrid technology, we start by a short explanation of this high-throughput technology for detecting physical bindings between proteins (see section 2.7.1). Furthermore, we describe two large data sets used in this work to derive functional modules, i.e. CORE and the von Mering data set (see sections 2.7.2 and 2.7.3), and two smaller data sets that focus on modules in signalling pathways (see sections 2.7.4 and 2.7.5).

2.7.1 Yeast two-hybrid (Y2H) technology

The yeast two-hybrid (Y2H) technology is based on the transcription activator GAL4, and its characteristic modular domain structure consisting of a DNA binding domain (BD) and transcription activation domain (AD) (Ito, et al., 2002). In the two-hybrid assay. two fusion proteins are created; the protein termed "bait". which is linked to the GAL4 binding domain. and its potential binding partner termed "prey", which is fused to the GAL4 activation domain. If the bait and prey proteins interact, their BD and AD (see Figure 9 on page 39) will combine to form a functional transcriptional activator (TA). This TA will then activate the transcription of reporter genes that are integrated in the region downstream of the GAL4 binding sites. A reporter gene is a gene whose protein product can be easily detected and measured. and the amount of its expression can be used as an indicator of interaction between the protein of interest (the bait) and its potential partner.

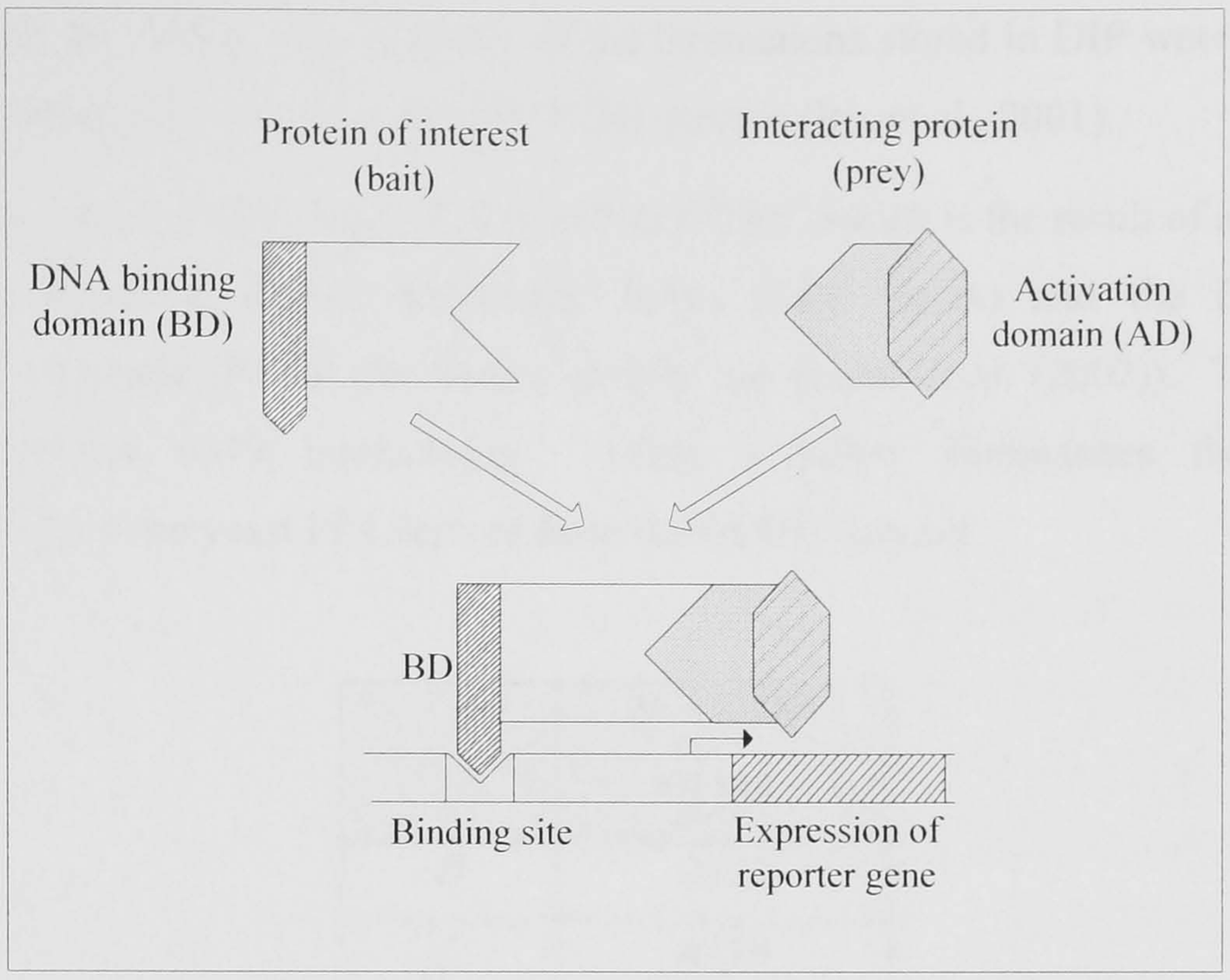


Figure 9: Basic mechanism of Yeast two-hybrid technology

Another commonly used method for high-throughput identification of protein-protein interactions is a combination of protein-complex purification followed by mass spectrometry (MS) (Mann, et al., 2001). One important distinction between Y2H and MS is that transient interactions are more often found by Y2H, whereas MS is more reliable in identifying stable interactions (Titz, et al., 2004).

However, there is no high-throughput method which is able to identify all protein-protein interactions. Many interactions get lost, and we will here mention some sources of this error. Sometimes, false negatives can be caused by the weak interactions within complexes that require cooperative action to be stabilised and generate a two-hybrid signal (Titz, et al., 2004). False positives are considered as a more serious problem than false negatives. The major reason for false positives in Y2H data is that true interactions that occur in the Y2H system may never occur *in vivo* because the proteins are for example expressed in different cell types. The number of false positives can be reduced to some extent by conducting additional assays and using computational methods.

2.7.2 Yeast CORE data set

The Database of Interacting Proteins (DIP: http://dip.doe-mbi.ucla.edu) is a database that stores and organises experimentally determined protein-protein interactions

(Xenarios, et al., 2000). The majority of the interactions stored in DIP were identified with high-throughput yeast two-hybrid (Y2H) screens (Ito, et al., 2001).

There is the subset of DIP-YEAST, denoted as CORE, which is the result of assessment with the Expression Profile Reliability Index (ERP Index) and the Paralogous Verification Method (PVM) (for further details, see Deane et al. (2002)). The CORE subset contained 6379 interactions. Table 1 below summarises the general characteristics of the yeast PIN derived from the CORE data set.

	Yeast PIN (CORE)
N	2231
L	6379
N_L	96
$\bar{k}$	4.8
C	0.34

Table 1: Network characteristics for the yeast PIN derived from CORE. N denotes the total number of proteins in the data set. L is the total number of interactions (195 self-interactions are excluded here), the number of proteins connected to the largest hub is denoted by N_L, the average degree is denoted by $\bar{k}$ and the average clustering coefficient is denoted by C.

Some of the large-scale properties of the network derived from the whole CORE data set are analysed in Chapter 6. The experiments in Chapters 6-9 are also based on this data set. The same data set is also used as basis for deriving different subsets which were further analysed in Chapter 5. A description of those subsets can be found in sections 2.7.4 and 2.7.5.

2.7.3 Protein-protein interaction data from von Mering

In a study presented by von Mering et al. (2002), a quality assessment of large-scale data sets of protein-protein interactions in yeast was performed. A critical evaluation of the accuracy of high-throughput data is needed, because of the high rate of false interactions in these data sets. In (von Mering, et al., 2002), data sets from yeast two-hybrid (Y2H) systems, protein complex purification techniques that rely on mass-spectroscopy (TAP and HMS-PCI), correlated mRNA expression profiles, genetic

interactions, and *in silico* interaction predictions were analysed. As stated further in this study, each of these methods can be used to predict protein interactions, even though their goals are slightly different. While the main purpose with yeast two-hybrid and mass spectrometry is to identify physical binding between pairs of proteins, the remaining of the mentioned methods are mainly focused on predicting functional associations, which in many cases also requires physical binding (von Mering, et al., 2002).

The authors integrated about 80 000 interactions between yeast proteins and found that only 2455 were supported by more than one method. This low overlap between sets of protein interactions obtained from different methods may be due to the high fraction of false positives, but may also be caused by the difficulties for some methods to capture certain types of interactions. All interactions are classified by the level of confidence (low, medium, high), based on the evidence that supports them. In our study, we have used the interaction set with high level of confidence, meaning that all interactions are confirmed by several methods. We will refer to this data set as "von Mering". The data set contains 2455 interactions between 988 proteins. Table 2 below summarises the general characteristics of this yeast PIN.

	Yeast PIN (von Mering)
N	988
L	2455
N_L	51
$\bar{k}$	5.0
C	0.55

Table 2: Network characteristics for the yeast PIN derived from von Mering. For explanation of row labels, see Table 1 on page 40.

We use this data set for comparison with the CORE data. Because this data set includes other types of interactions besides the experimentally determined ones, it is interesting to analyse similarities and differences between modules derived from different data sets. The experiments based on this data set can be found in Chapters 7, 8, and 9.

2.7.4 Yeast filamentation network

Budding yeast, when subjected to nitrogen starvation, undergoes a morphological change from yeast growth form to invasive filamentous form. The filamentation network responds to this stimulus by changes in metabolism, cell cycle progression, polarized budding, cell-cell adhesion and invasion. We choose to analyse this network since this process is regulated by several conserved pathways that are well studied (Prinz, et al., 2004; Rives and Galitski, 2003). Furthermore, it is hypothesized in (Rives and Galitski, 2003) that the modular abstraction of the filamentation network allows us to relate the observed filamentous cell properties with the activation or repression of specific biological processes.

Most prominent among the pathways involved in the filamentous growth are mitogen-activated protein-kinase (MAPK) and the cAMP/PKA pathway (Rupp, et al., 1999). Since the virulence of many human fungal pathogens is controlled by highly related signalling processes, an understanding of filamentous-form growth and directed disruption of it can reduce their ability to cause disease.

In (Rives and Galitski, 2003), a yeast filamentation network was clustered, based on shortest path distance between the nodes, to identify modular structure. The network was identified from the database in combination with a literature search. After elimination of the proteins having no interactions with other filamentation proteins, this filamentation network contained 70 proteins, 61 of which were present in the CORE data set. These 61 proteins and the interactions among them that were extracted from the global CORE data set constitute the part of the yeast filamentation network used in the study that is described in Chapter 5. The input graph with protein-protein interactions for yeast filamentation proteins is exemplified in Appendix A.

2.7.5 Yeast signalling network

The second network that was derived from the CORE data set for the purpose of module extraction is the yeast signalling network. The yeast signalling network is an intensively studied system that contains well-defined modules of signalling pathways. The list of proteins belonging to the signalling protein category was obtained from MIPS (http://mips.gsf.de/genre/proj/yeast/). This category contains 133 proteins, 89 of which remained after removing proteins that were not present in CORE and those not having any interactions with other signalling proteins. Interactions among the

remaining 89 proteins were extracted from the global CORE set of interactions. This subset is also used in the study that is described in Chapter 5.

Chapter 3

Related research

In this chapter, we will focus on the related methods for identifying modular structures. We introduce this chapter with a short overview of the common purposes for deriving such structures, to provide an overall picture of the related work in this area. Various approaches for module identification are described in sections 3.1-3.3.

The modular organisation of cellular networks offers testable hypotheses that lead to biological insights. We could discern three major lines of research in previous work that are successfully utilising modular organisation to learn about different aspects – from functions of individual proteins to the regulation of cellular processes.

The first important aspect is that genes in a given module are hypothesised to be functionally related. For instance, modules from PIN, exemplified in (Qi and Ge, 2006), contain proteins involved in common functions such as RNA polyadenylation and chromatin remodelling, indicating that there is a strong correspondence between network topology and functionality. This further suggests that uncharacterised genes or proteins belonging to modules could be functionally annotated accordingly. Several methods have used Gene Ontology to predict function of hypothetical proteins from the protein-protein interaction graphs (Deng, et al., 2004; Karaoz, et al., 2004).

The second important aspect is that module structures provide information about the underlying regulatory mechanisms of the cell. Large-scale gene expression data is commonly used for this purpose. Based on gene expression data, Segal et al., (2003) predicted regulatory modules consisting of regulators and their potential target genes, along with the conditions under which the regulatory relationships are relevant. By inspection of the transcriptional changes of potential target genes as a result of a

disruption of regulator functions, the regulatory roles of several previously uncharacterised transcription factors were identified (Segal, et al., 2003). The application of this in module-level analysis of cancer data has later revealed a global view of the shared and unique modules that underlie human cancer (Segal, et al., 2004). The third important aspect is that studies of inter-modular connections have confirmed that cellular processes are coordinated events (Petti and Church, 2005). For example, connections between glycolysis and lipid metabolism modules revealed their transcriptional coordination (Petti and Church, 2005).

Hence, there are many reasons for developing module-identifying methods and to study connections between them in order to understand the organisation that underlies cellular functionality. Several methods have been proposed to identify functional modules on the basis of the topology of the interaction network (Giot, et al., 2003; Girvan and Newman, 2002; Rives and Galitski, 2003; Spirin and Mirny, 2003). In the following sections, we review some of the methods that are related to our work.

3.1 Clustering coefficients for analysis of protein networks

Several studies have used various types of clustering coefficients to find dense sub-graphs in PINs. Although the clustering coefficient is a good measure of the density of interactions in a protein interaction sub-graph, it is strongly dependent on the size of the sub-graph. This makes it difficult to use clustering coefficient values to discern sub-graphs for which the density is statistically significant.

Spirin and Mirny (2003) elaborated on this problem by starting from each sub-graph with n proteins and m interactions and computing the probability of obtaining more than m interactions among the same set of proteins in a random graph. They observed that the majority of cliques of size four or greater are statistically significant in PINs compared with random graphs. This statistical significance of cliques points to the importance of the functions that the cliques carry, because they have emerged under selection. Such enrichment in the number of cliques reveals essential modularity in the network structure, suggesting that many of these protein interactions are responsible for the formation of complexes and functional modules. More specifically, they used three different methods for identifying protein complexes from a PIN, where all methods were focused on identifying highly connected clusters. The PIN was constructed based on the protein interactions from MIPS database. The first method consisted of

identifying all cliques by complete enumeration. The largest clique that was found contained 14 proteins. The second method they used was clustering by using a so called superparamagnetic clustering algorithm. Finally, in the third method, the problem of finding densely connected clusters was formulated as an optimisation problem, i.e. finding the set of nodes k that maximise the function $C(n,k) = 2n/k(k-1)$. The right-hand side of this equation corresponds to the clustering coefficient as a measure of density. All discovered modules or complexes were related to MIPS functional annotation and they found that the majority of those identified belonged to the following four functional classes: RNA processing, transcriptional regulation. cell-cycle/cell-fate control, and protein transport.

Bader and Hogue (2003) developed an algorithm for detecting molecular complexes in PINs, called MCODE (Molecular COmplex DEtection). MCODE uses the node weighting scheme that is based on the clustering coefficient defined by Watts and Strogatz (1998). They used the notion of highest k-core of a graph. which is the most densely connected sub-graph. Based on this notion, they introduced the term core-clustering coefficient of a node i, and defined it as the density of the highest k-core of $N[i]$. The final weight assigned to a node is a product of the core-clustering coefficient and the highest k-core number, called k_{max}, of the $N[i]$. After the weighting step, proteins are sorted in decreasing order of their weights and the highest weighted protein is used to seed a complex. From the seed node, the algorithm moves outwards, including into the complex nodes with weights above a given percentage of the weights of the seed node (Bader and Hogue, 2003). The procedure is repeated with the next unexplored node with highest weight. At this stage, the algorithm does not allow any overlap between complexes. The post-processing step that allows a certain overlap may be chosen.

Bader and Hogue (2003) evaluated their complexes against two data sets of complexes, the set of complexes found in MIPS (also used in this work) and the data set from (Gavin, et al., 2002). To find optimal parameter settings for MCODE, they used 221 complexes from (Gavin, et al., 2002) to evaluate MCODE and found that predicted complexes matched 88 out of 221 Gavin complexes. Also, the same evaluation procedure was repeated for MIPS complexes, where MCODE predicted 166 complexes out of which only 52 matched MIPS complexes. Such low agreement between predicted and MIPS complexes may be attributed to incompleteness of the MIPS

complex catalogue, the incompleteness of the data set that MCODE was run on, or the possibility that human annotated definitions of complexes do not perfectly match with a density-based definition. Possible future work could, according to Bader and Hogue (2003), include other scoring functions that take into account functional annotation of the proteins.

Those two approaches described in this section have probably influenced our work mostly. Since both are purely topological methods, they seemed to offer space for improvement by including semantic information, which is the main direction in our work.

3.2 Clustering approaches based on graph theoretic properties

Protein complexes and modules may be found by clustering the protein interaction network with respect to some topological properties, such as mutual clustering coefficient (Poyatos and Hurst, 2004) or shortest path length (Rives and Galitski, 2003).

Rives and Galitski (2003) developed and evaluated a clustering method based on a shortest-path distance matrix between all pairs of proteins. Each distance was transformed into a value that the authors call "association", which corresponds to $1/d^2$, where d is the shortest-path distance. Hierarchical agglomerative average-linkage clustering was used to cluster the association matrix. The method was applied on a network of signalling proteins to identify the modular organisation of cellular networks controlling specific biological responses. The following assumptions were made in this work: 1) the shortest path between any two nodes is likely to represent a functional association and information transmission; 2) each node in a network has a unique profile of shortest-path distances that comprises all distances from the node to all other nodes in the network, and 3) proteins that belong to the same module are likely to have similar profiles of shortest-path associations. Rives and Galitski (2003) demonstrated that the protein clusters in the signalling network based on shortest-path association profiles represent modules of signalling pathways, where more than one module may exist within the pathway. Furthermore, they clustered the network of proteins that are associated to filamentation process in yeast. By applying the same method as for the previous network, they generated a clustering that reflects and extends current models of the filamentation process. We found some elements of this work to be arbitrary, even though the basic idea of using shortest-path association profiles to reveal modules is

very interesting. In our work, proposed in Chapter 5, we introduce the novel idea of generating profiles of semantic similarities as a feasible way to represent functional associations between proteins, in combination with mutual clustering coefficient profiles. To avoid arbitrary choice of cluster boundaries, which is determined by using visual means in (Rives and Galitski, 2003), we attempt to build in a validation procedure into the clustering step, where a cut-off is determined based on the agreement with functional annotation.

Another network clustering method based on the graph-theoretic properties of the nodes was proposed by Poyatos and Hurst (2004). In their study, the assumption was made that phylogenetic information could be used to verify putative interaction-based modules. They clustered proteins based on two properties; profiles of shortest-path associations, proposed in (Rives and Galitski, 2003), and profiles based on the mutual clustering coefficient measuring common neighbours of any two proteins, proposed by Goldberg and Roth (2003) and described in section 2.3.3 on page 23. Those two matrices were clustered separately with a standard hierarchical agglomerative average-linkage clustering algorithm. Matrices were then subjected to an overlap procedure, where the highest overlap between clusters based on different properties served as a guide for extracting modules according to the number of branches present in the clustering tree. It was found that the average maximal overlap between different clusterings was significantly high, that is equal to or greater than 0.8. This result indicates that both topological measures generate similar clusters, which in turn limits the added value of generating modular structures based on both measures. In our study, described in Chapter 5, we attempt to add a novel element to this method by combining one of the topological properties, profiles based on mutual clustering coefficient, with annotation-based semantic similarity profiles.

In a study by Pereira-Leal et al. (2004), graph clustering is performed after transforming the graph consisting of edges that connect nodes into its associated line graph in which edges now represent nodes and nodes represent edges. The advantage of such a transformation is, according to the authors, that it generates the higher-order local neighbourhood of interaction, which in turn results in a more highly structured graph compared to the original graph. An observed fivefold increase of the average clustering coefficient after transformation of the original graph was used to illustrate this. They used an algorithm for clustering by graph flow simulation, called TRIBE-MCL

(Enright. et al.. 2002). (where MCL denotes Markov Clustering) to cluster the interaction network and find functional modules. These clusters were then transformed back to the original protein-protein graph for subsequent validation. Validation of the obtained clusters was performed by assessing the consistency of the protein classifications within each cluster. For this purpose, they used a measure called redundancy, which reflects the homogeneity of classification within a cluster. Three different classification schemes were used in the evaluation: regulatory and metabolic pathway schemes from KEGG (Kyoto Encyclopedia of Genes and Genomes). an automated functional classification scheme from GeneQuiz, and a cellular localization scheme from MIPS/CYGD.

3.3 Modular decomposition

Modular decomposition is an algorithmic method that can be used to define the organisation of protein-protein interactions as a hierarchy of nested modules (Gagneur. et al., 2004). This approach emphasizes the identification of protein complexes and the way they share components. In (Gagneur. et al., 2004) the higher order logical organisation between the proteins in protein-protein interaction networks is represented by using a graph-theoretic definition of modules. According to this definition, a module consists of a set of nodes that have the same neighbours outside the module. The modular decomposition of a graph is a tree of modules and proteins. with the modules as internal nodes and the proteins as leaves.

The elements of a module can be replaced by a representative node, called "quotient". because all nodes in a module share the same neighbours outside the module (Gagneur, et al., 2004). Quotients can be iterated and captured in a tree where each node represents a module at a specific hierarchical level and the leaves represent the individual proteins. The root of the tree represents the whole network and the labels of the nodes show the relationships of that node's children. There are three types of modules classified according to the logical relationships between the proteins they include. They are called prime, series and parallel, which has been found to correspond to biological strategies of protein reuse (Gagneur. et al.. 2004). In a series module. all proteins, or modules are grouped together and can bind directly to each other (logical "and"). while a parallel module consists of proteins that are alternative binding partners to their common neighbours (logical "xor"). In a prime module some proteins, but not all. can bind to each other. From the structural point of view. proteins in a parallel

module share the same or overlapping binding sites, while proteins in a series module are likely to have non-overlapping sites (Gagneur, et al., 2004). This interpretation that the authors propose seems logical, but this should also be verified in order to confirm the biological plausibility of this method. Furthermore, the authors state that it would be tempting to decompose the networks based on Y2H interactions, but they could not come up with an appropriate functional interpretation applicable to Y2H data. This method has been applied on large PINs based on PCP (protein-complex purification) experiments and it clearly helped structuring the network according to the proposed higher-order logical structures. However, the method itself should be combined with other strategies that are appropriate for the purpose of the system of study (Gagneur, et al., 2004).

Chapter 4

Semantic similarity measures for predicting protein interactions

Determination of protein-protein interactions is fundamental for our understanding of the molecular machinery of the cell. The function of a protein is defined to a great extent by its interaction with other proteins (Winters and Day, 2003) and can be seen as its position within the cellular interaction network (Salwinski and Eisenberg, 2003). Thus, the inference of a protein's interacting partners is a vital step towards the identification of its role within a cell. The recent emergence of high-throughput technologies for the detection of protein-protein interactions, such as yeast two-hybrid screens (Ito, et al., 2001), has resulted in a rapid accumulation of protein–protein interaction data. However, the data is fairly noisy, and differences between the interacting proteins identified in high-throughput experiments and those generated with traditional small-scale experiments point at the need for computational approaches for data validation. Many studies have assessed the quality of the high-throughput data (Mrowka, et al., 2001; von Mering, et al., 2002) and confirmed that some of the data sets contain many false positives.

In the study described in this chapter, we use semantic similarity measures to assess the quality of interactions. We investigate if protein-protein interactions can be predicted by using the similarity between the proteins based on their ontology terms. We also provide a comparison between three of the most widely used semantic similarity measures. According to our first research question (see section 1.2), our aim is to compare the predictive power of GO-based semantic similarity measures, when applied

on data sets with varying degrees of clustering. An appropriate semantic similarity measure will be chosen to calculate weights, in terms of similarity between proteins. If these measures prove to have good predicted power, they can be useful for validating protein-protein interactions data. A possible application would be to predict the functionality of new proteins. Information on protein interactions was downloaded from the Database of Interacting Proteins (DIP) (see Chapter 2), which contains experimentally determined interactions between proteins in *S. cerevisiae*.

4.1 Materials and methods

In this section we describe the way of discerning three subsets of CORE that have different properties in the network. We also present the evaluation procedure that is used to compare the predictive power of semantic similarity measures.

4.1.1 Identification of CORE subsets

Three different subsets of proteins were derived from CORE. The results from protein-protein interaction predictions for each of the sets were then compared to investigate how different characteristics of the chosen subsets may affect the predictive power of the similarity measures.

As the purpose of the study performed in this chapter is to test how the degree of the clustering of different subsets (as an indicator of a certain architecture of the network) may affect the predictive power of the semantic similarity measures, we needed to set an appropriate threshold on the clustering coefficient value, to be able to discern between more densely interconnected regions and the regions that are sparsely interconnected. In previous work (van Noort, et al., 2004), a network architecture of gene coexpression in *S. cerevisiae* was studied, because it is regarded as a general indicator of protein involvement in the same biological processes. The clustering coefficient (C') of the coexpression network with significantly high correlation (for further details, see (van Noort, et al., 2004)) was 0.6, with an average shortest path of 4. Thus, this network showed the properties of a small-world, scale-free network that is characteristic for intracellular networks in which the nodes that are connected to each other are often involved in the same processes. We have chosen the threshold $c > 0.6$ to create the data set that follows this architecture, under the assumption that semantic similarity-based predictions will perform better in such network, where the interaction between proteins should reflect their involvement in the same process. Similarly, using

data sets derived from lower clustering coefficient values should lead to inclusion of random connections that do not fit in the small-world network. which will probably affect the performance of semantic similarity-based predictions.

The first subset is called CORE-CC06 and it contains highly interconnected proteins that are likely to form modules. As clustering coefficient values reflect the degree of interconnectedness among the proteins in a module, we used this measure to rank the proteins accordingly. The nodes with clustering coefficient values that exceed 0.6 and their neighbouring proteins were selected to build a subset containing 318 proteins. It was shown in (Lubovac, et al., 2005) that the more interconnected nodes are in a module, the higher is the average semantic similarity within the module. Consequently, we expect all three tested measures to perform well in predicting protein-protein interactions in this data set.

The second subset, CORE-CC05, contains randomly selected proteins from CORE that have a clustering coefficient $c < 0.6$. This set also contains 318 proteins, as CORE-CC06, and was created for comparison with CORE-CC06.

The third subset, CORE-CDC28, contains the protein CDC28 and all of its neighbours, 96 proteins in total. CDC28 is one of five different cyclin-dependent protein kinases (CDKs) in yeast (Mendenhall and Hodge. 1998) and has the highest connectivity of all CORE proteins. CDC28 is fundamental in the control of the main events of the yeast cell cycle (Mendenhall and Hodge, 1998) and therefore interacts with many proteins. However, because it acts as a hub, i.e., holds together several functionally related clusters, it does not necessarily have high sequence or semantic similarity with its neighbours. Therefore, the results from protein-protein interaction predictions for the proteins in the CORE-CDC28 data set were expected to contain a considerable number of false interactions.

In this study, we calculated semantic similarity with tree different information theoretic measures originally proposed by Lin (1998), Jiang and Conrath (1998), and Resnik (1999), respectively (see section 2.6.1). We investigated if it is possible to set a threshold on semantic similarity values that will separate interactions from non-interactions. The pairwise similarities between all pairs of proteins in each data set were calculated, irrespective of whether the proteins shared a connection in the interaction network or not. Semantic similarity calculations are based on the annotation derived using the sub-ontology covering molecular function.

4.1.2 Evaluation

For each of the three CORE subsets described in section 4.1.1, pairwise semantic similarity was calculated between all pairs of proteins, using all three semantic similarity measures. This resulted in three matrices for each subset, where each entry denotes semantic similarity between a pair of proteins. Each matrix was then converted into a binary interaction matrix, where 1 stands for an interaction and 0 stands for no interaction. Each entry was assigned 1 if the semantic similarity between the pair of proteins was above a specific threshold value; otherwise it was assigned the value 0. We refer to these matrices as ss_{PIN}-matrices. These matrices were compared with the original interaction matrix, referred to as the PIN-matrix, created from CORE, where 1 denotes an interaction and 0 denotes a non-interaction. The PIN-matrix was then used to calculate sensitivity and specificity of the interactions in the ss_{PIN}-matrices.

The threshold value of the minimum semantic similarity required between a pair of proteins to get score 1 in the ss_{PIN}-matrix was varied to see which threshold value would predict the experimentally determined interactions with the highest specificity and sensitivity. The interval within which the threshold value was varied was determined by manually inspecting the results from the calculations of the semantic similarities between the proteins. Different intervals were used for the three different semantic similarity measures.

We compared each of the ss_{PIN}-matrices with the PIN-matrix by varying the semantic similarity thresholds to determine the predictive power of each semantic similarity measure. Sensitivity and specificity scores are used in the evaluation. Sensitivity is the probability that the method will correctly identify positives and it is defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

where TP is the number of true positives and FN the number of false negatives. Specificity indicates how well the method is able to reject negatives, and is defined as:

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

where TN is the number of true negatives and FP the number of false positives.

The sensitivity and specificity of a predictive method varies with the chosen threshold. Receiver Operating Characteristic (ROC) curves are used to display the range of sensitivities and specificities of a prediction method. ROC curves were created for each subset of data and each measure of semantic similarity. A ROC curve is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) and shows the trade-off between sensitivity and specificity of a prediction method, i.e. an increase in sensitivity will result in a decrease in specificity. The area under the ROC curve is a measure of the accuracy of the prediction method. An area of 1 signifies a perfect prediction method whereas an area of 0.5 represents a method for which the predictive power is no better than random guessing (Tape, 2005). A traditional academic point system, as described in (Tape, 2005), can be used as a rough guide for classifying the accuracy of a prediction method ([0.9-1.0]-Excellent (A); [0.8-0.9]-Good (B); [0.7-0.8]-Fair (C); [0.6-0.7]-Poor (D); [0.5-0.6]-Fail (E), where the numbers in squared brackets denote the area under the ROC curve).

Furthermore, in this study we use sensitivity multiplied with sensitivity to determine suitable cut-off points and compare different measures. In previous work, this measure has been used for evaluation of cancer diagnosis efficiency (Mori, et al., 2003; Obermann, et al., 2005). The same quantity has also been used to determine suitable cut-off points, as in Peulen et al. (1998). There are other measures for this purpose, such as the positive and negative prediction power, but they are not evaluated here.

4.2 Results

As mentioned earlier, we evaluated the ability of three semantic similarity measures to predict protein-protein interactions at different threshold values, by calculating the sensitivity and specificity of the predictions. Since the sensitivity of the predictions is typically highest at lower threshold values while the specificity is highest at higher threshold values, it is difficult to set the ultimate threshold on both measures. In this section, the three semantic similarity measures will, for convenience, be referred to as the Lin, Resnik and Jiang-Conrath measure, respectively.

All of the three measures were found to perform with fairly high specificity and sensitivity. There is a slight difference in predictive power between the measures, but as expected, all three measures show significantly better prediction power in the three data sets CORE-CC06, and CORE-CC05 compared to the data set CORE-CDC28 (see Table 3 on page 56).

	CORE-CC06	CORE-CC05	CORE-CDC28
Lin	0.80	0.85	0.43
Resnik	0.79	0.89	0.43
Jiang-Conrath	0.76	0.89	0.47

Table 3: Summary of the predictive power of Lin, Resnik and Jiang-Conrath

Each entry in Table 3 above denotes the highest value of the product between specificity and sensitivity for respective data set.

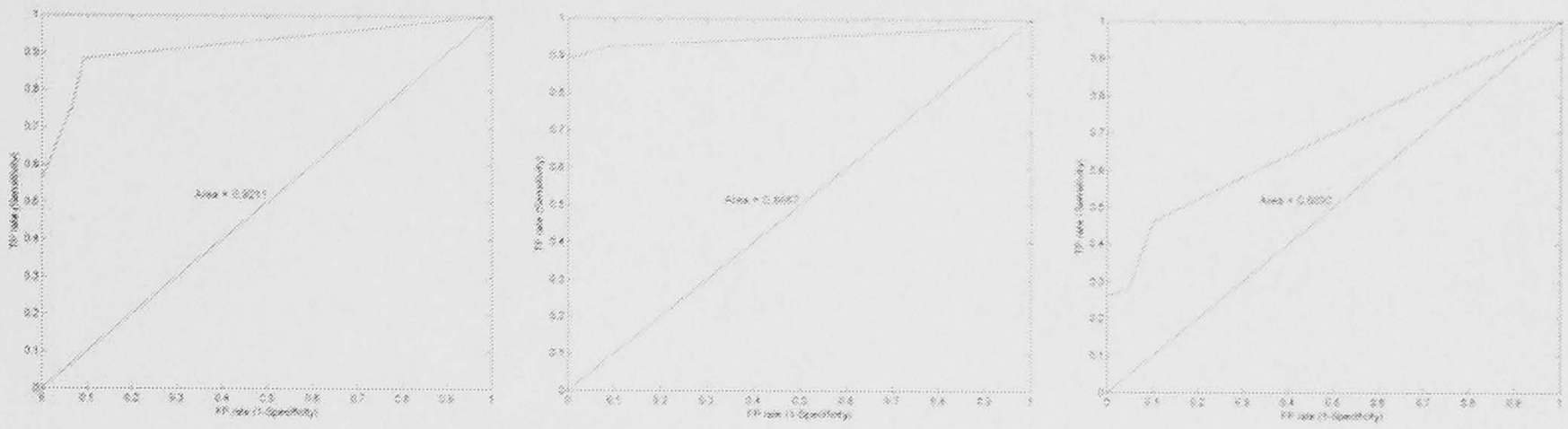
4.2.1 Evaluation of the Lin measure

The Lin measure performed best of the three measures in one of the three data sets, namely, CORE-CC06. On the other hand, for the CORE-CDC28 data set Lin was the measure that predicted the protein-protein interactions with the lowest accuracy. Nevertheless, this measure has its advantages over the Resnik measure. For instance, it takes all functions of the proteins into account, i.e., the average of all specific terms that are assigned to the protein and their parents' functions, whereas the Resnik measure only uses the information content of the shared parents (see section 2.6.1). The best results produced by applying the Lin measure are shown in Table 4 below and the corresponding ROC curves are shown in Figure 10 on page 57. According to the guide for classification of predictive power (see section 4.1.2), predictions for the data set CORE-CC06, fall into category A (excellent accuracy). The predictions for CORE-CC05 are classified as category B (good accuracy), and for the CORE-CDC28 the results fall into category D, i.e. poor accuracy.

Data set	Threshold	*Specificity*	*Sensitivity*	*Specificity* * *Sensitivity*
CORE-CC06	0.5	0.91	0.88	0.80
CORE-CC05	1.0	0.93	0.92	0.85
CORE-CDC28	0.1	0.76	0.56	0.43

Table 4: Results from the Lin measure applied on the three data sets

In Figure 10 below, the ROC curves shows the results from using the Lin measure with the threshold value varied between 0 and 1. The threshold value for each data set was set using the highest product of specificity and sensitivity as a guide. The best result is obtained for CORE-CC06 (area = 0.92), as shown in Figure 10a, followed by the ROC curve for data set CORE-CC05 in Figure 10b (area = 0.87), and the lowest area under the ROC curve is obtained for CORE-CDC28 (area = 0.69), and is shown in Figure 10c.



(a) **(b)** **(c)**

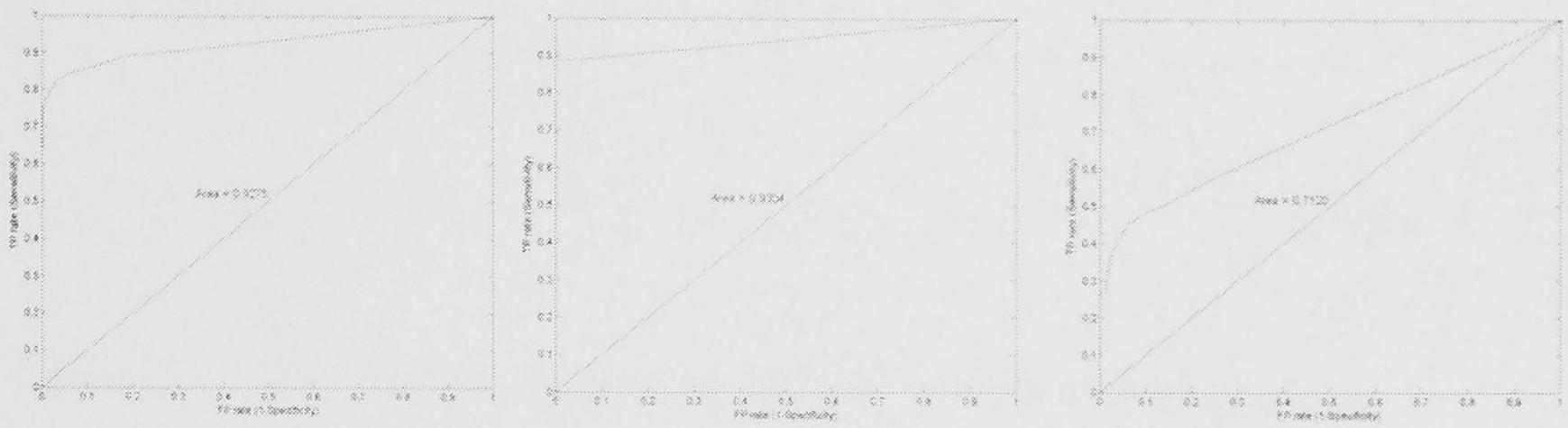
Figure 10: ROC curves showing accuracy using the Lin measure

4.2.2 Evaluation of the Resnik measure

The Resnik measure performed best in predicting the protein-protein interactions for one of the three data sets, CORE-CC05, together with Jiang-Conrath. In previous work by Lord et al. (2003), it was illustrated that the Resnik measure, based on GO molecular function between proteins, showed the strongest correlation with sequence similarity. Also this measure showed the worst predictive performance on CORE-CDC28 data set. The best results from using the semantic similarity measure by Resnik on the three data sets are shown in Table 5 on page 58. The ROC curves for the results obtained when using the semantic similarity measure by Resnik to predict protein-protein interactions in all three data sets are shown in Figure 11 on page 58. For data set CORE-CC06, area = 0.93, and is shown in Figure 11a. The same size of the area under the ROC curve was obtained for CORE-CC05, and the smallest area, 0.71, was obtained for CORE-CDC28 (see Figure 11c). ROC curves are obtained by varying the semantic similarity threshold value between 0 and 10. Hence, the predictions for two of the data sets, CORE-CC06 and CORE-CC05, fall into category A, i.e. excellent accuracy. The predictions for data set CORE-CDC28 fall into category C, which indicates fair accuracy.

Data set	Threshold	*Specificity*	*Sensitivity*	*Specificity * Sensitivity*
CORE-CC06	2.0	0.94	0.85	0.79
CORE-CC05	8.0	1.00	0.89	0.89
CORE-CDC28	1.0	0.75	0.58	0.43

Table 5: Results from the Resnik measure applied on the three data sets



(a) **(b)** **(c)**

Figure 11: ROC curves showing accuracy using the Resnik measure

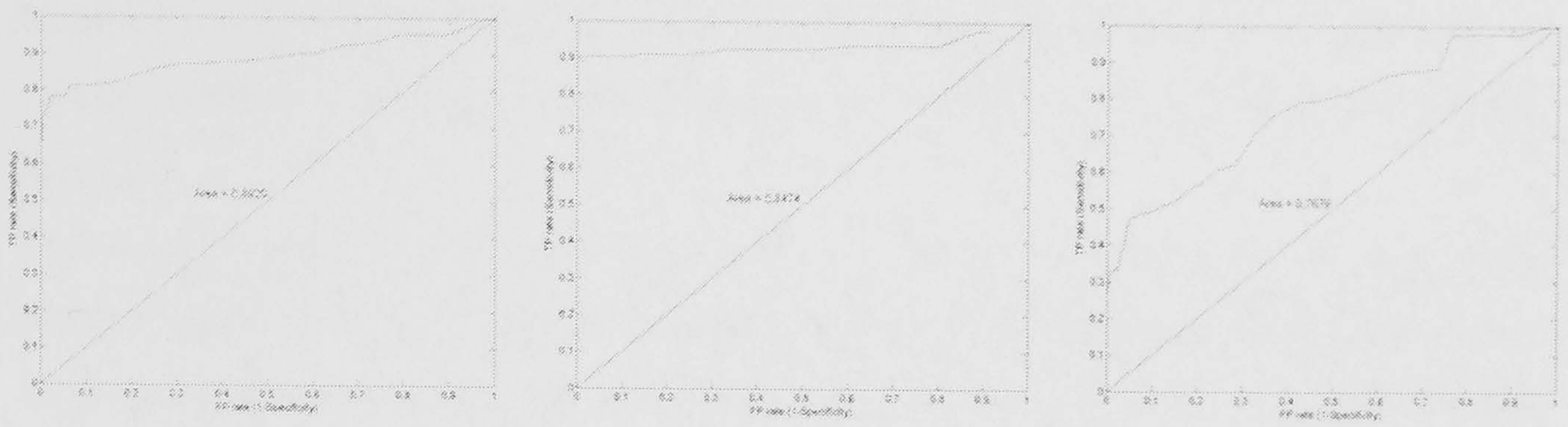
4.2.3 Evaluation of the Jiang-Conrath measure

The third and final measure, Jiang-Conrath, showed the worst predictive accuracy of the data set CORE-CC06, but the best for the CORE-CDC28 data set (see Table 6 below and Figure 12 on page 59). This suggests that the properties of the data set CORE-CDC28 do not matter as much to the Jiang-Conrath measure as to the other two measures. Furthermore, this measure has the same advantage as the Lin measure over the Resnik measure, that is, it includes all functions of the proteins and not only those of the shared parents.

Data set	Threshold	*Specificity*	*Sensitivity*	*Specificity * Sensitivity*
CORE-CC06	18.00	0.98	0.78	0.76
CORE-CC05	34.00	1.00	0.89	0.89
CORE-CDC28	12.00	0.67	0.70	0.47

Table 6: Results from the Jiang-Conrath measure applied on the three data sets

ROC curves of the protein-protein interaction predictions are generated using the Jiang-Conrath measure with the threshold value varied between 0 and 35 (see Figure 12 below). For this semantic similarity measure, predictions for the data sets, CORE-CC06 (area = 0.89) and CORE-CC05 (area = 0.85) belong to category B (good accuracy), and for CORE-CDC28 with area = 0.77, the results fall into category C (fair accuracy).



(a) **(b)** **(c)**

Figure 12: ROC curves showing accuracy using the Jiang-Conrath measure

4.3 Summary and conclusions

All three measures of semantic similarity seem to be able to predict protein-protein interactions with good specificity and sensitivity and thus, in combination with the Gene Ontology, seem to be good tools for such predictions. The differences between the results from different measures for predictions in the same data set are small. However, the difference between the performances of the measures in different data sets is considerably large. It is clear that the accuracy of predictions is much higher for two of the data sets, CORE-CC06 and CORE-CC05, compared to the third data set, CORE-CDC28. It was expected that all methods would predict the interactions in the CORE-CC06 data set with higher accuracy than the interactions in CORE-CC05, but that was only the case with the Lin and Jiang-Conrath measures. The Resnik measure predicts the interactions in CORE-CC05 with highest accuracy. The measure by Lin performed best on the data set that potentially involves modules, as it contains proteins that are highly interconnected, which is reflected by the value of their clustering coefficient. This is one of the reasons for our decision to use this semantic similarity measure in continuing studies to generate weights between protein-protein interactions. A later study (Posse, et al., 2006) also supports our decision, as they show that the correlation between the other two GO aspects, biological process and cellular component, and sequence similarity, is highest using the Lin semantic similarity, based on cellular

component and biological process sub-ontologies (for the later one, the correlation is equal when using the Jiang-Conrath measure).

For the last data set, CORE-CDC28, the measure by Jiang and Conrath generated the best results followed by the measure by Resnik and then the one by Lin. This may suggest that the measure by Jiang and Conrath is less affected by the properties of the data set it is applied on than the other two measures, even though the difference in accuracy is quite large between the data sets that the measure performed the best and the worst on.

In conclusion, the difference between ROC curves for different measures was so small that any of the semantic similarity measures may have been used in further experiments. A predominant reason for our final choice of the Lin measure (along with already mentioned reasons) was that it uses the probability of specific terms rather than the probability of the shared parent.

In (Lord, et al., 2003), it was shown that semantic similarity based on the molecular function aspect of GO showed strongest correlation to sequence similarity of all of the GO aspects. This was the main reason for choosing this aspect in our study. Not all proteins that interact have similar functions or sequence and therefore it would be interesting to take the other two aspects into consideration, especially the biological process aspect. The accuracy of the predictions would probably increase if a combination of two or three aspects would be used.

Chapter 5

A cluster overlap approach based on topological and domain knowledge

Various methods of network clustering have been applied to reveal modular organisation in protein-protein interaction networks (Pereira-Leal, et al., 2004; Poyatos and Hurst, 2004; Rives and Galitski, 2003). However, those methods have mostly been based on structural properties of the network, such as shortest path distance for example. In this chapter, we introduce a method that uses both topological and domain knowledge based on semantic function similarity derived from GO annotation. The method is based on clustering of two types of protein profiles, one based on mutual neighbours information, calculated with mutual clustering coefficient, and the other one based on GO semantic similarity. This novel combination of clusterings is then merged to a single modular structure. The proposed method has been applied to known modular networks, to test if it is able to recover known modules, and to evaluate the interconnectedness between modules.

In previous work, an approach based on topological properties has been applied to generate overlapping functional modules (Pereira-Leal, et al., 2004). Here, we propose that the use of functional annotation, in combination with topology, can add additional knowledge about functional modules.

5.1 Materials and methods

The method proposed in this study consists of a four-step procedure for generating and merging clusterings to derive a modular structure, thereby identifying the key modules of a protein interaction network. The method is based on combining the topology information, in terms of a mutual neighbours profile for each protein with its corresponding profile based on GO semantic similarity). In step one, matrices are generated to store topological information about the network structure and functional information about the proteins. In step two, a clustering algorithm is applied using each matrix as input, which results in sets of protein clusterings based on topological and functional information. In step three, individual clusterings which show a high degree of overlap are identified. and in the fourth step, the chosen clusterings are merged and a modular structure extracted.

We here choose to analyse two networks, a filamentation and signalling network, since they involve processes that are regulated by several conserved pathways, and are well studied (Prinz, et al., 2004; Rives and Galitski, 2003). A more detailed description of those networks may be found in sections 2.7.4 and 2.7.5.

In the following sections, we describe the input data and details regarding the algorithms that make up the approach.

5.1.1 Overview of the method

To derive modules from protein interaction networks, we use a four-step procedure, which is described in this section and illustrated in Figure 13 on page 63. The most important difference between our method and the one proposed in (Poyatos and Hurst, 2004) is that our method uses domain knowledge (the *SS* matrix) in combination with topological properties (the *CC* matrix), whereas (Poyatos and Hurst, 2004) only used topological properties.

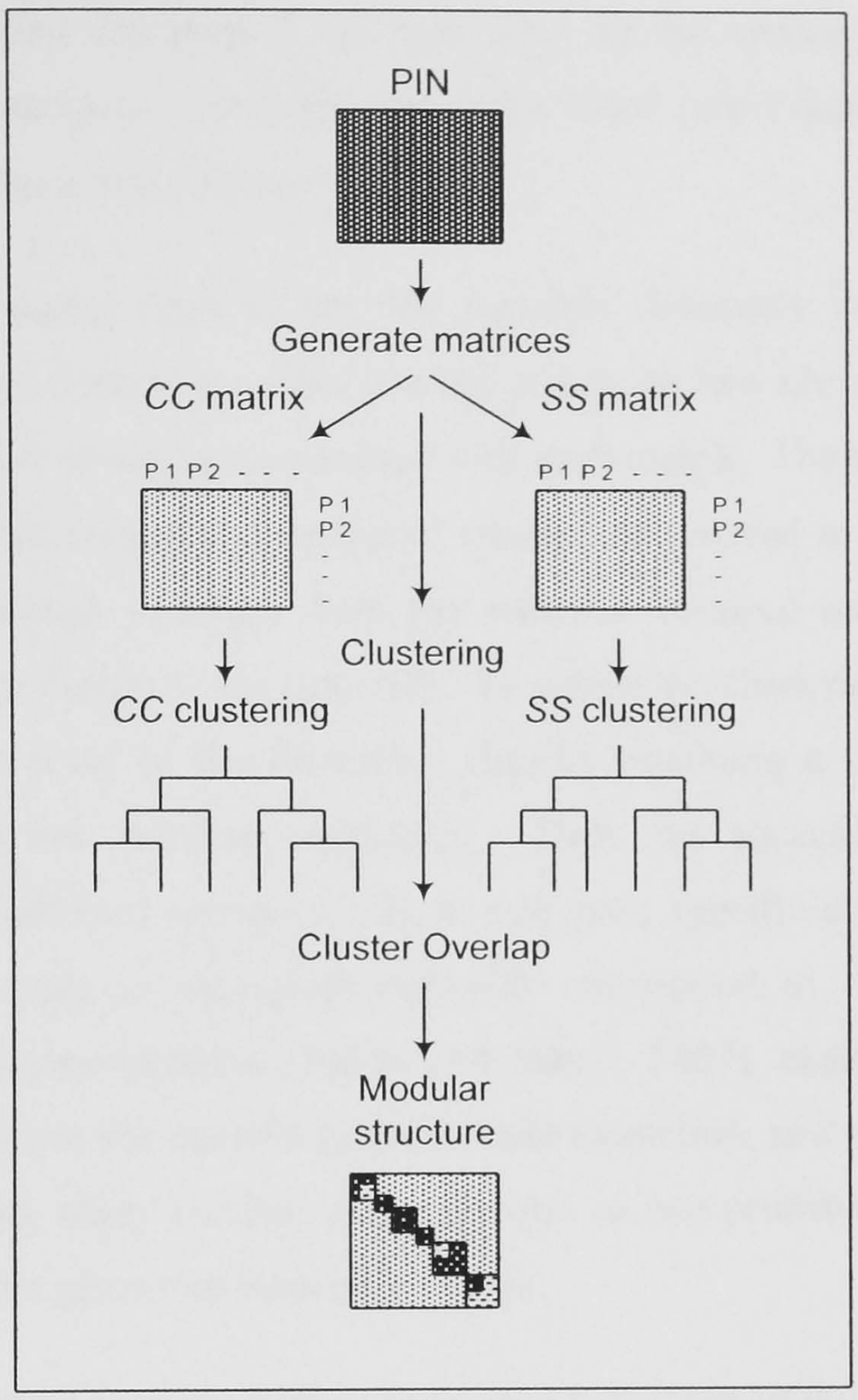


Figure 13: Deriving a modular structure based on cluster overlap

- Step 1 is to generate matrices (*CC* and *SS*, see Figure 13 above) based on the chosen properties of the proteins that form the network. The choice of properties is vital in this process, since the matrices constitute a basis for later module extraction. Those properties should be able to reflect the modularity. The presence of a modular topology in the PIN could be manifested in the fact that proteins within a module are likely to interact more frequently with each other than with proteins outside the module, which is reflected in a high value of global *C*. Alternatively, modularity could also imply that proteins within a module share similar processes or functions to perform a common activity. In one of the matrices (*CC*), each entry contains the mutual clustering coefficient value for the actual interaction. The mutual clustering coefficient reflects the local topology by measuring the common number of neighbours of any two proteins in the interaction graph. For more details on the mutual clustering coefficient, see

section 2.3.3 and Equation 2 on page 23. In the second matrix (*SS*). each interaction is assigned a semantic similarity value (see Equation 5 on page 35) between the interacting proteins.

- Step 2 is to cluster each of the two matrices separately using agglomerative average-linkage clustering. This process results in two sets of clusterings with varying numbers of clusters associated with each matrix. The clusterings obtained from the mutual clustering coefficient matrix are referred to as *CC* clusterings, whereas clusterings obtained from the semantic distance matrix are called *SS* clusterings (see Figure 13 on page 63). To obtain the clusterings, we first cut the tree at a high level of the hierarchy, thereby obtaining a clustering with few branches and low minimal similarity. Then, we successively increase the similarity cut-off (see section 5.1.2). to get more specific clusters. Since most relevant processes in biological networks correspond to the mesoscale, i.e., involve 5–25 genes/proteins (Spirin and Mirny, 2003), clusterings that diverge from this scale are not interesting for module extraction, and we therefore discard trees containing many clusters with only one or two proteins. Typically, 10-15 clusterings were generated from each matrix.

- Step 3 is to calculate the average overlap between each of the *CC* clusterings and all those obtained from the *SS* clusterings. This step is referred to as cluster overlap in Figure 13. The purpose is to find the pair of clusterings with best average overlap (see section 5.1.3), which will serve as a basis for deriving the modular structure. We also calculate the overlap ratio (see section 5.1.3), to get an indication of whether a cluster in one clustering overlaps with one or more clusters in the other clusterings. The choice of best candidate for a generating modular structure is based on the greatest average overlap *O* between *CC* and *SS* clusterings with as low overlap ratio *OR* as possible.

- Step 4 is to extract a modular structure (see Figure 13) from the pairs of clusterings chosen in step 3. For each cluster in the chosen *CC* clustering, a new module is generated by merging this cluster and the cluster with the greatest overlap in the chosen *SS* clustering. In (Poyatos and Hurst, 2004), the modular structure was instead extracted by keeping in each *CC*-based module only those proteins which also appeared in the module from the corresponding clustering

(which, in their case is based on shortest path distance). Each approach has its advantages. The approach by Poyatos and Hurst (2004) may result in more topologically robust modules, for example. Under the assumption that the functional similarity, which is the basis for deriving one of the matrices, is a vital part of module extraction, we propose that it is more advantageous to merge clusters, rather than to intersect. By extracting a modular structure based on the intersection between clusters, we would get the modules of proteins which share a similar neighbours profile and have similar functions. However, when merging two clusters, we also include the proteins that belong to the *SS*-based cluster but not to the *CC*-based cluster, since those proteins share similar functions even if they do not necessarily share a similar neighbours profile with the rest of the cluster. Likewise, proteins belonging to the *CC*-based cluster but not to the *SS*-based cluster are considered valuable to include in a module since this might give us a possibility to hypothesise about their new alternative functions.

It is important to highlight that the choice of the candidate for extracting a modular structure is based on the best matches between clusterings, which reduces the risk for introducing many false positives. When a pair of clusterings with a considerably large overlap is used for module extraction, the majority of the proteins in the generated modules do indeed belong to the intersection set between two clusters, but in some modules there are a few additional proteins belonging to the difference sets between clusters, which are interesting in further analysis. It is also important to note that the *CC* clustering is used as a template to guide the overlap procedure, i.e. we start from clusters based on topological properties to find the best overlap matches with functional properties. This is because the clustering coefficient is known to be a robust measure for protein network analysis, and we do not want modular extraction to be governed only by the knowledge-based clustering. Instead, we let the *SS* clustering support and influence the modular structure to some extent, rather than to guide the whole process.

5.1.2 Hierarchical clustering

The protein interaction graph is used to generate a *CC* and *SS* matrix for the two analysed networks described in sections 2.7.4 and 2.7.5. We store mutual clustering coefficient values (see Equation 2 on page 23) for all node pairs in the *CC* matrix, while the *SS* matrix stores semantic similarity values (see Equation 5 on page 35) for all node pairs. The two matrices, *CC* and *SS*, are then clustered separately with Hierarchical

Clustering Explorer (Seo and Schneiderman, 2002). Dendrograms are generated with the agglomerative average linkage clustering method UPGMA (Unweighted Pair-Group Method using Arithmetic average), using Pearson correlation as distance metric. A feature provided by the clustering software, called dynamic query control, is used to identify clusterings at different levels in the hierarchy. To achieve different clusterings or subgroups, the minimum similarity bar may be pulled down, which splits the dendrogram into two, three, four, etc. clusters. Any pair of objects in the clusters below the similarity bar are more similar to each other than the minimum similarity threshold specified by the bar. Objects that are distant from the cluster are removed. Tighter clusters appear as the minimum similarity threshold increases. Domain knowledge stating that the most relevant processes in biological networks involve 5–25 genes/proteins is used as a further guide in obtaining clusterings, which resulted in clusterings containing 4-15 clusters for each matrix.

5.1.3 Cluster overlap

We use the overlap algorithm, originally proposed in (Ihmels, et al., 2002), and later applied by Poyatos and Hurst (2004) for the identification of cluster overlaps based on the shortest path distance and mutual clustering coefficient. The overlap between two different clusters C_i and C_j is defined as:

$$Ol_{ij} = \frac{|C_i \cap C_j|}{\sqrt{|C_i| \times |C_j|}} \tag{10}$$

When modular structures are derived (see section 5.1.1), the average overlap is used to decide which clusterings should be chosen for deriving the modular structure. The average overlap is defined as (Poyatos and Hurst, 2004):

$$O = \frac{1}{n_{cc}} \sum_{i=1}^{n_{cc}} \max\left\{Ol_{i,j\ j=1..n_{ss}}\right\} \tag{11}$$

where n_{cc} and n_{ss} denote the number of clusters in the CC and SS clusterings, and Ol is the overlap defined in Equation 10 above.

An additional measure, used for increased reliability of the chosen clustering is the overlap ratio (OR), defined as (Poyatos and Hurst, 2004):

$$OR = \frac{1}{n_{cc}} \sum_{i=1}^{n_{cc}} \left| \left\{ Ol_{i,j \; j=1...n_{ss}} = \max \left\{ Ol_{i,j \; j=1...n_{ss}} \right\} \right\} \right| \tag{12}$$

The value of OR should be as low as possible. If $OR = 1$, there is only one cluster in the SS clustering that maximally overlaps with a cluster in the CC clustering, while $OR > 1$ indicates that there is more than one maximally overlapping cluster.

5.2 Results

5.2.1 Filamentation network

In previous work by Rives and Galitski (2003), the yeast filamentation network was clustered based on the profiles of shortest path distances between all proteins in the network. The underlying hypothesis in (Rives and Galitski, 2003) is that module co-members are likely to have similar shortest path distance profiles. However, as pointed out in (Watts and Strogatz, 1998), the shortest path distance method is less efficient than the mutual clustering coefficient in identifying modular structure in so called small-world networks. We therefore start by deriving a clustering based solely on the mutual clustering coefficient approach to compare the obtained structure with the modular structure based on shortest path distances presented in (Watts and Strogatz, 1998). Thereafter, we extract a modular structure according to the four-step procedure described in section 5.1.1. This structure, based on combined properties, is also compared to the clustering based solely on topological properties of the network.

Topology-based approach

We start by clustering the yeast filamentation network based on mutual clustering coefficient values (see Equation 2 on page 23). Before applying a combined approach, we generated a clustering based solely on topology to see what additional knowledge we can gain by introducing knowledge-based clustering. A symmetrical matrix of 61 proteins of the yeast filamentation network was clustered based on the mutual clustering coefficient. The resulting dendrogram is shown in
Figure 14 on page 68 and is highly similar to the one generated in (Rives and Galitski, 2003).

There are at least two conserved signalling cascades regulating filamentous growth (Lengeler, et al., 2000): the cAMP-PKA (cyclic adenosine monophosphate/protein kinase A) pathway, and the filamentous MAP kinase cascade (fMAPK). The topology-

based approach identified both as modules. Other identified modules that also agree with those identified in (Rives and Galitski, 2003) are CDC28, polarity, SNF and HMLα and Ras (see Figure 14 below). BMH, which was part of the polarity module in Rives and Galitski's results, here emerged as a separate module.

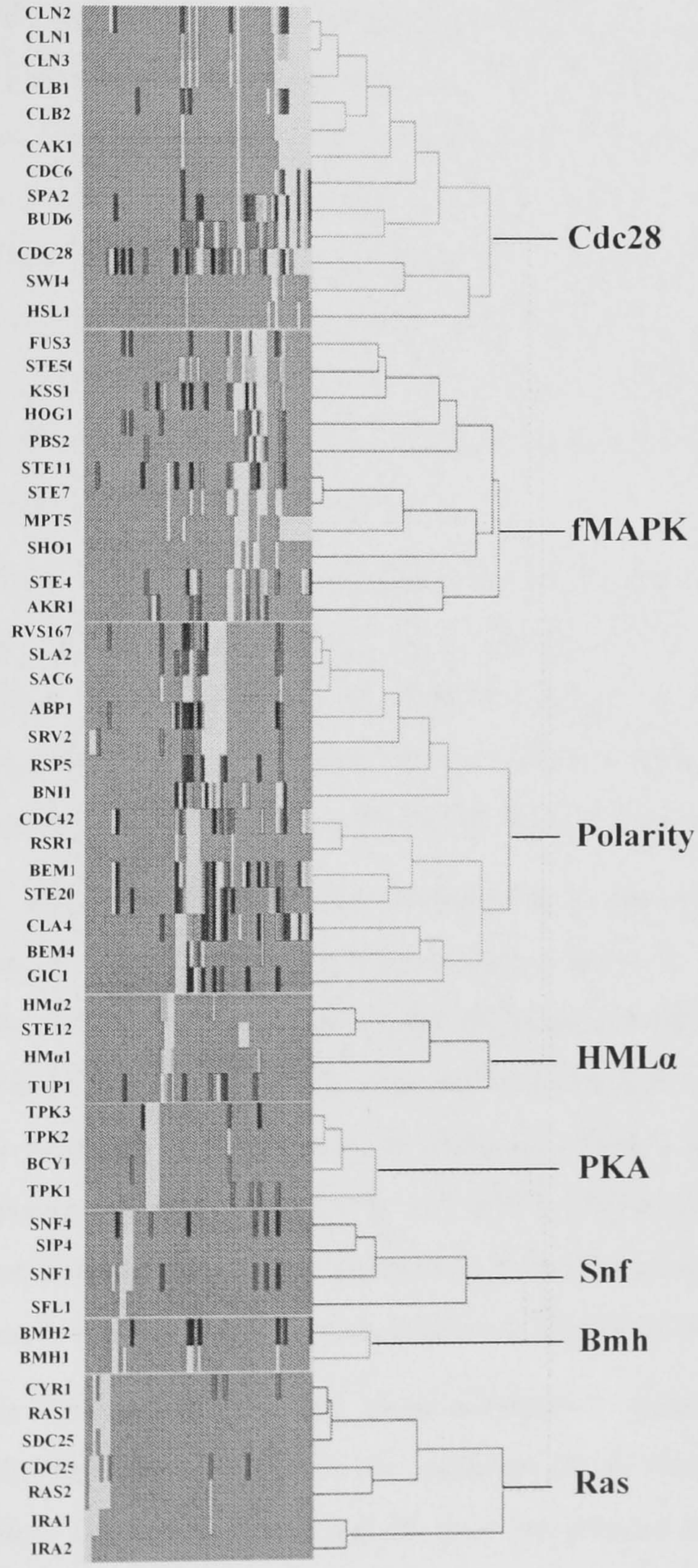


Figure 14: Clustering of the yeast filamentation network

One of the differences between the modules obtained in this work and the modules from (Rives and Galitski, 2003) is that Rives and Galitski identified part of the HOG (High Osmolarity Glycerol response) pathway as a separate module, whereas it is here clustered with other proteins belonging to the fMAPK module. The HOG pathway has two redundant input branches that activate a common target, Pbs2. Besides being a MEKK (an activator of MAPK) for the HOG pathway, Pbs2 also acts as the scaffold for a branch of the pathway that shares components with the fMAPK pathway. The shared component between fMAPK and HOG is MEKK kinase Ste11. Sho1, which is a member of the same cluster, is required for activation of Pbs2 via Ste11 (Posas and Saito, 1997). It is obvious that fMAPK and HOG share components to such an extent that the approach based on the mutual clustering coefficient places both pathways into the same module. It is proposed in (Rives and Galitski, 2003) that more than one module can exist within the pathway. Here, we also found an example of the existence of more than one pathway within the module.

Another protein whose identified module membership differs from the one proposed in (Rives and Galitski, 2003) is Ste12. Ste12 is a transcription factor involved in the kinase cascade and it is associated with the fMAPK module in (Rives and Galitski, 2003). In our study, Ste12 is clustered together with proteins associated with cell-type specification and mating, and belongs to so the called HMLα module (see

Figure 14 on page 68). The Ste12 protein mediates transcriptional induction of cell type-specific genes by binding to a DNA sequence referred to as the pheromone-responsive element (PRE), which is present in the upstream control region of several *a*- and α-specific genes (Yuan, et al., 1993), such as Hmlα1 and Hmlα2, which are also present in the same module. This motivates the presence of Ste12 in the HML module, but since it is a common target of the fMAPK and mating (HMLα) pathways, it should preferably belong to both modules. This problem is possible to cope with by allowing overlapping proteins in the modular structure, which we introduce in the next section.

The PKA module corresponds to the protein-kinase-A pathway that regulates filamentous growth. The Bcy1 protein is an inhibitor of all three cAMP-dependant protein kinases Tpk1, Tpk2, and Tpk3, and all four are present in the same module. Deletion of Bcy1 enhances filamentous growth (Pan and Heitman, 1999). The SNF module corresponds to glucose-control proteins that play an important role in the filamentation process, since the glucose depletion causes invasive growth (Cullen and

Sprague, 2000). The CDC28 module, highly similar to the one identified in (Rives and Galitski, 2003), emerged in the present study as well. Finally, the Akr1 protein did not fall into any of the clusters in the work by Rives and Galitski (2003). In our work, Akr1 is placed in the fMAPK module, which seems to be biologically plausible since it interacts with Ste4 (Kao, et al., 1996), which is one of the major components of the fMAPK module.

The method described in this section, as well as the method described in (Rives and Galitski, 2003), suffers from the disadvantage that it results in a set of disjoint clusters, which is not biologically plausible since proteins may function in several modules. Our combined method attempts to cope with this problem by allowing overlapping proteins. Furthermore, we also use the functional information about the interactions, along with the topological information, to obtain functional modules where proteins should have consistent functional annotation. We should strive for a highly consistent functional annotation of the proteins within a module, since it indicates their common involvement in biological processes and their common function (Pereira-Leal, et al., 2004).

Combination of topology- and semantic-based approaches

After applying an approach based solely on a topological property of the network, in this case the mutual clustering coefficient, we tested a combined approach, where knowledge in terms of GO annotation is used in addition to the topological measure. The purpose is to investigate whether this combination can add some additional knowledge to the modular structure. Two aspects of GO have been tested, namely molecular function and biological process. Since the topological approach turned out to be very robust in previous work (Goldberg and Roth, 2003; Poyatos and Hurst, 2004) and able to derive modules here that are highly similar to the modules confirmed in previous work, we used the matrix based on this property as a basis or template for module extraction and compared it with other matrices obtained from different GO-aspects.

Extracting modular structure is based on the overlap between different clusterings and it is done according to the four-step procedure explained in overview in section 5.1.1. When testing the knowledge-based approach, we first consider the functional aspect of GO. For comparison, we also generated random clusterings, to get an indication of the significance in the overlap between different clusterings. The best average overlap ($O \approx 0.6$) was found between a *CC*-based clustering containing nine clusters and a *SS*-

based clustering containing 12 clusters. Comparing the same *SS* clustering with randomly generated clusterings resulted in significantly lower average overlap values ($p = 9.94 \cdot 10^{-7}$). We could therefore conclude that the overlap with clusters based on functional knowledge is significantly higher than the overlap with randomly generated clusters, which is expected.

We also compared the average overlap values between *CC* clusterings and *SS* clusterings based on GO annotation regarding biological process and molecular function. The average overlap values for process-based clusterings were slightly higher than those obtained from clusterings based on molecular function. However, this difference is not statistically significant ($p \approx 0.55$). Since both knowledge-based sources performed equally well in combination with the mutual clustering approach. modular structures were derived based on both combinations. For details on the module extraction procedure, see section 5.1.1.

The modular structure derived from the best overlapping *CC* clustering and an *SS* clustering based on molecular function is illustrated in Figure 15a on page 72 while the corresponding structure with an *SS* clustering based on biological process information is illustrated in Figure 15b. We will refer to those modules as functionally informed and process-informed modules, respectively. Figure 15 shows the semantic similarity between all proteins in the analysed filamentation network. The original network contains 61 proteins, but some proteins were assigned to multiple modules, based on GO molecular function, which results in a matrix with 74x74 entries (see Figure 15a). The matrix in Figure 15b, based on GO biological process annotation, contains 70x70 entries. In Figure 15, numbers along x and y axes represent proteins, which are ordered according to module membership. White entries represent protein pairs with semantic similarity $ss > 0.5$. There are nine functionally informed modules shown in Figure 15a. Furthermore, in Figure 15a, the functional homogeneity within the modules is apparent, whereas proteins within modules in Figure 15b seem to have more disparate GO-terms reflecting biological processes. More specifically, functionally informed matrix contains 143 entries between proteins that belongs to modules with $ss > 0.5$, which is 44% of the total number of possible pairs of module members. The corresponding fraction of protein pairs with $ss > 0.5$ for process-informed modules is 28% (99 entries with $ss > 0.5$ out of 351 in total). Functional homogeneity is further evaluated with so called redundancy value (see section 5.3 and Equation 13 on page 75).

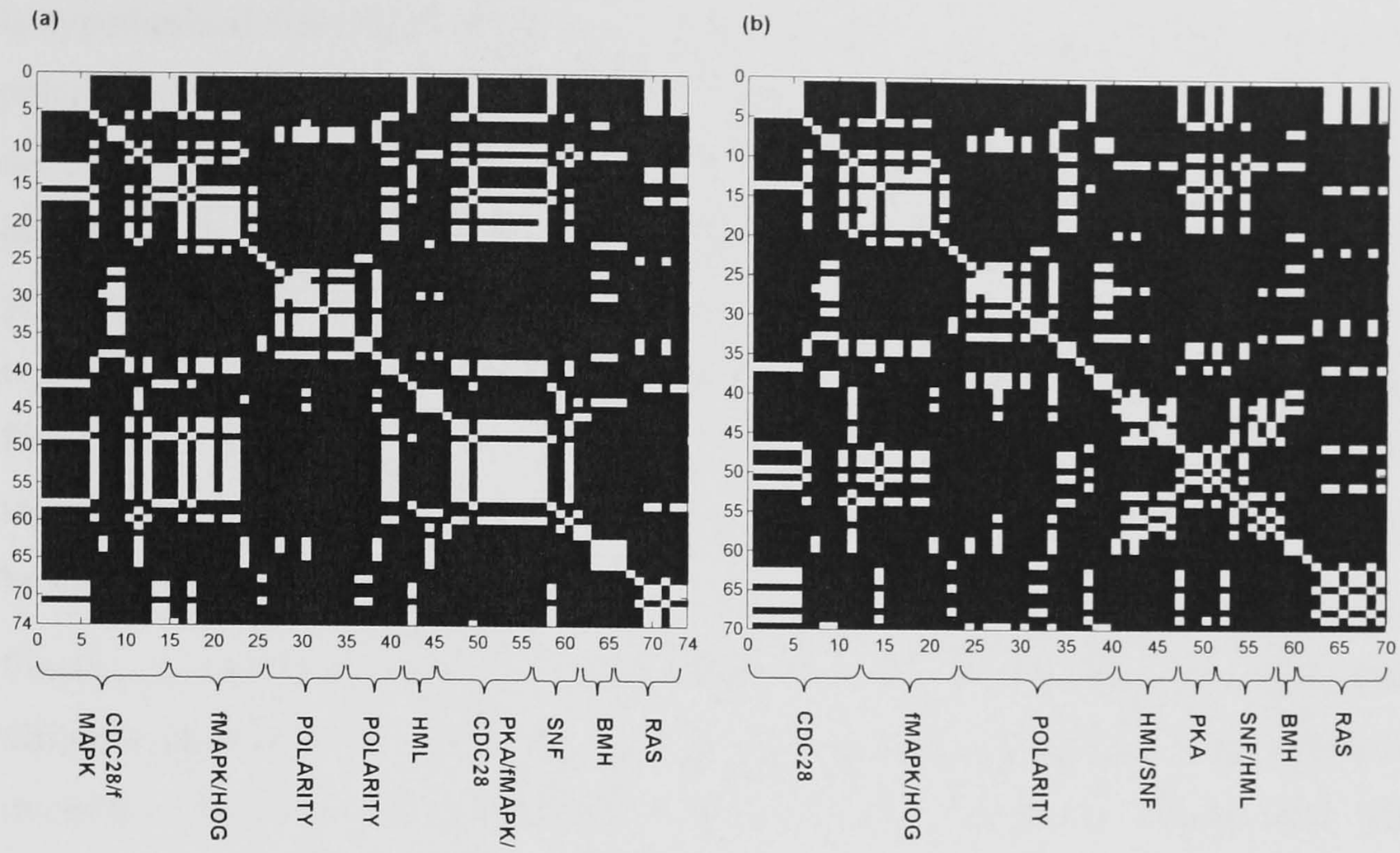


Figure 15: (a) Functionally informed modules (b) Process-informed modules

We further investigated what additional knowledge we can gain by generating a modular structure from a combination of properties. Concrete examples include the Bud6 and Spa2 proteins, which appear in the polarity module when using the combined approach, but not when using topology alone. Also, since each module derived by the combined approach represents the union of two clusters with the best overlap, several modules may share the same proteins. Here, the Bud6 and Spa2 proteins are found both in the CDC28/fMAPK module and the polarity module (this module is expanded with those two members compared to the corresponding topological counterpart in Figure 14 on page 68). In previous work by Rives and Galitski (2003), those proteins were identified as parts of the fMAPK module. One of the proteins, Bud6, is also identified in (Rives and Galitski, 2003) as an intermodule connector between the polarity and fMAPK modules. As suggested in (Rives and Galitski, 2003), such connectors are critical for intermodule communication and may even function in both the modules they connect. This is why it is particularly interesting that this protein could be found in both modules, because the presence of the same protein in different modules may give some clues about the point of crosstalk between pathways.

Another illustrative example is the Gic1 protein, which only appeared in the polarity module when using topology alone, but emerged both in the functionally informed Ras and polarity modules after the cluster merging procedure. In (Kozminski, et al., 2003) it

is hypothesized that Ras-family proteins interact directly with Cdc42 (that belong to the polarity module) to promote the establishment of cell polarity. Ras2, for example, appears to have a Cdc42-dependent role in polarized growth. Ras2 also plays an important role in the induction of the polarized morphogenesis required for filamentous growth, by signalling via the Cdc42/MAPK kinase module (Mosch, et al., 1999; Mosch, et al., 1996). However, the mechanism by which Ras2 signals to Cdc42 is unknown. Since Gic1, like Cdc42, belongs to the Rho-family GTPases, it is tempting to speculate that Gic1 also has a mediating role between Ras- and Rho-family GTPases, which may be important for filamentous growth.

Finally, a module containing a combination of the proteins associated with four different pathways, which have connected roles, occurred among functionally informed modules. We call it the PKA/fMAPK/CDC28 module (see Figure 15a on page 72). This module contains PKA-associated proteins (Tpk1-3 and Bcy1), fMAPK-associated proteins (Cla4/Ste20, which are largely redundant, and Ste7), Cdc28-associated proteins (Cdc28, Hsl1 and Cak1), and finally there is one protein, Snf1, responding to carbon starvation. PKA and fMAPK are parallel pathways and are both required for filamentous growth. It is also known that Cdc28-associated proteins are involved in the altered cell-cycle progression of the filamentous-form cells, but there is no evidence that Cdc28 directly links to the fMAPK cascade or the PKA signalling pathway involved in filamentous growth. In (Edgington, et al., 1999), two lines of evidence are described, which indicate that Cdc28 is a controlling factor in the decision to grow in the filamentous or yeast form.

Process-informed modules (see Figure 15b on page 72) show great similarity with functionally informed modules, but also some important differences that may contribute to additional knowledge. One of the differences concerns the Ras module. Compared to both its topological and functionally informed counterpart, this module contains three additional proteins that overlap with other modules and seem to be vital for intermodule communication. For example, Ste4, which can also be found in the fMAPK module, constitutes a component of an intermodule path between fMAPK and Ras. Srv2 is another protein that is placed both in the Ras and Polarity modules, and has previously been identified as a mediator of intermodule communication between those two modules (Rives and Galitski, 2003).

5.2.2 Modular structure of the yeast signalling network

We also used the yeast signalling system described in section 2.7.5 to test our combined approach. Since our cluster overlap procedure implies that a protein may be present in multiple functional modules, the clustering of 89 proteins resulted in a modular structure containing 150 entries. The best average overlap ($O \approx 0.5$) occurred between a *CC* clustering containing 11 clusters and an *SS* clustering containing 12 clusters. Each protein is on average present in 1.7 modules.

The result of the module extraction procedure is presented in Figure 16 below. In the figure, functionally informed modules of the yeast signalling network are indicated by numbered curly brackets. The large symmetrical matrix illustrates 150 proteins of the MIPS-database signalling category. Each number along x and y axes represents a protein and they are ordered according to their module membership. White entries in the matrix represent protein pairs with functional $ss > 0.5$. Columns to the right represent different signalling pathways (F/M: Filamentation/Mating; R: Ras; P: Polarity; H: HOG).

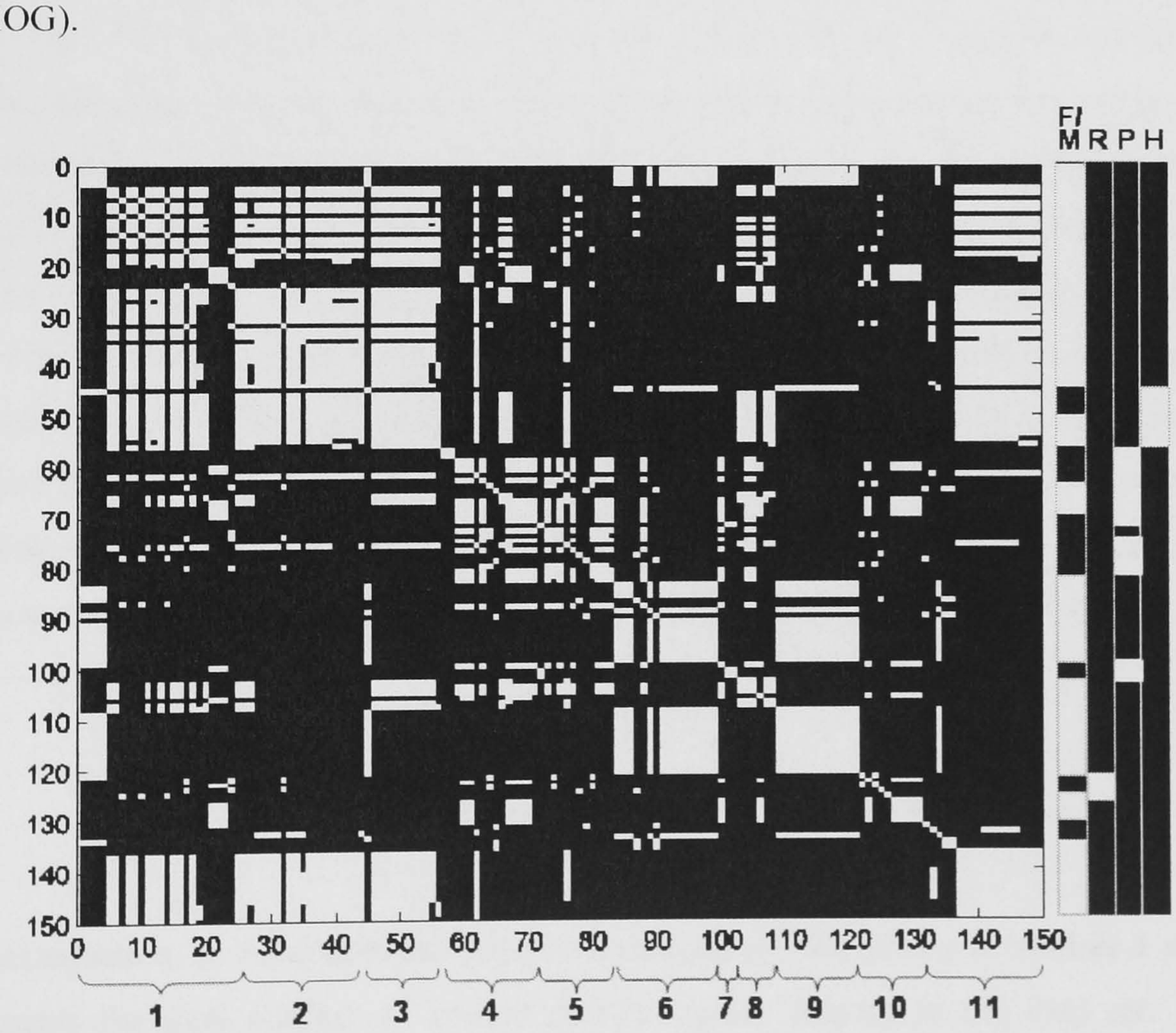


Figure 16: Functionally informed modules of the yeast signalling network

In Figure 16, we can notice that the Ras pathway-associated proteins form a single cluster. The majority of the HOG-pathway proteins are also placed in a separate module. However, some pathways, like the Filamentation/Mating pathway, stretch over several clusters, which reflects the existence of more than one module within the pathway. This has previously been shown in (Rives and Galitski, 2003). However, unlike the approach used in (Rives and Galitski, 2003), our approach produces overlapping modules. Some proteins, such as Ste11, are present in both the HOG and filamentation/mating modules, which is interesting since this is a shared MAPKKK cascade component and a point of crosstalk between the HOG and filamentation pathways.

5.3 Evaluating the functional homogeneity of the obtained modules

Besides biological evaluation, we have applied an independent computational evaluation method based on another most widely used source of annotation, obtained from Munich Information Centre on Protein Sequences (MIPS). Protein categories from MIPS/CYGD (Mewes et al., 2002) are used to measure the functional consistency within individual module. For more information about MIPS, see section 2.6.3. This classification schema is hierarchical and provides several levels of granularity in the classification, from general classes at the top level of the hierarchy to more specific classes lower down. For evaluation of filamentation network, we used second level of hierarchy, while we used even more specific third level for evaluation of modules obtained from signalling network, since network consists of the proteins belonging to the first level of MIPS category "Intracellular signalling". Each module is assigned a value that reflects the homogeneity of the protein classification within that module. The measure called redundancy R_i for module i (Pereira-Leal et al, 2004) is defined as:

$$R_i = 1 - \left(\frac{-\left(\sum_{j=1}^{n} p_j \log_2 p_j \right)}{\log_2 n} \right) \quad (13)$$

In this equation, p_j represents the relative frequency of the class j in module i and n represents the total number of unique protein classes present in the data set. The numerator represents the information content in bits given by Shannon's entropy, and the maximum entropy for the module i is used as a normalisation factor (Pereira-Leal et al., 2004). The value of R ranges between 0 and 1, where modules with proteins that

have highly consistent classifications will receive high value, whereas those with highly inconsistent classifications will receive low value. For comparison, we calculated average redundancy value for all combinations of clusterings, not only the one with the greatest average overlap that represents modular structure.

We evaluated homogeneity of the protein clusters within each module with redundancy value R_i for module i. Along with average overlap and overlap ratio measure, the redundancy measure could also be integrated in the module-identifying procedure and serve as a natural step in identification of functionally homogeneous modules. Here, it is only used in evaluation purpose. For each pair of clusterings, average redundancy $\overline{R}$ is calculated. Average redundancy is average value of the redundancies of each obtained module, where each module is represented by the union of the cluster based on mutual clustering profiles information and its knowledge-based counterpart with maximal overlap. All modular structures do not necessary have highest redundancy value, but they had significantly higher consistency of functional categories within modules that the structures derived from the clusterings with significantly lower overlap.

We performed Wilcoxon signed rank test to evaluate if there is significant difference between functional homogeneity of modular structures based on GO molecular function and the corresponding structures based on GO biological process. Results from filamentation network show that $\overline{R}$ values for functionally informed modules are significantly higher than those for process-informed modules ($p = 0.002$). We also compared the functional homogeneity of the derived modules with randomly derived modules. As expected, both functionally informed modules and process-informed modules show significantly higher functional consistency than the randomly derived modules ($p = 0.9 \cdot 10^{-3}$ and $p = 0.003$). The results are consistent for signalling network as well, where functionally informed modules show higher consistency in functional annotation than the process-informed modules ($p = 0.8 \cdot 10^{-4}$).

5.4 Summary and conclusions

We have proposed a method for deriving modular structures by merging clusters originating from dendrograms based on the mutual clustering coefficient (*CC*) and semantic similarity (*SS*). Clusters are merged by union, i.e., the derived modules include proteins that belong to the *SS*-based cluster but not to the *CC*-based cluster,

since proteins may share similar function even if they do not have similar neighbourhood profiles. Similarly, proteins belonging to the *CC*-based cluster but not the *SS*-based cluster are considered valuable to include in a module since this can provide clues to functions which were previously unknown and therefore not present in the current annotation.

One of the distinct advantages of this method is that it may generate overlapping modules of the interaction networks, which implies that a protein may be present in multiple functional modules. Many clustering approaches cannot place objects in multiple clusters, which is not biologically realistic, since proteins may participate in multiple cellular processes and pathways. Results from the analysis of both the filamentation and signalling networks imply that our method generates modular structures, where proteins that are assigned to more than one module in several cases play important roles in intermodule communication. Ste11 is an example protein that is present in both the HOG and filamentation/mating modules, which is of particularly great interest since this is a shared MAPKKK cascade component and a point of crosstalk between HOG and filamentation pathways.

This work contains important extensions/improvements of the previous work that were considered as valuable contributions.

Extracting a modular structure based on both functional knowledge stored in the annotations and a topological property (mutual neighbours profiles) is advantageous over extracting modules based solely on topological properties. In hierarchical clustering, the choice of appropriate cut-off is often arbitrary or based on visual analysis. An advantage of the proposed approach is that the cut-off is chosen based on the best overlap with the domain knowledge.

Under the assumption that functional similarity, which is the basis for deriving one of the matrices, is a vital part of module extraction, we propose that it is more advantageous to merge clusters into a union of both clusters, rather than to intersect, and thereby eliminate the possibility to potentially uncover new functions of the proteins.
We also performed validation of the functional homogeneity of the derived modular structure. It is done by comparing redundancy values that score functional enrichment of annotation terms within different modular structures based on different overlaps, and not necessarily maximal overlaps. The comparison with random clustering was done to

indicate the expected outcome of a clear functional enrichment of annotation terms in functional clusterings compared to their random counterparts.

As also indicated earlier, the redundancy value may be incorporated into the module-identifying procedure, by choosing modular structure based on the weighted scheme that maximizes the sum of the redundancy value and the overlap value.

A possible continuation of this work is to integrate cellular component information into the clustering approach, to strengthen the evidence for biological validity of the obtained modules. Some proteins, although able to interact, are never in close proximity to each other in the cell. This constraint could be solved by introducing cellular component information, which is incorporated in our later work described in Chapter 7.

Chapter 6

Identifying modules in PINs with semantic similarity weighted network measures

When a high value of average clustering coefficient C is used as a signature for potential modularity, it is important to realise that this measure is defined solely on topological grounds. Topology-based approaches may not be sufficient, especially when dealing with error-prone data. In this work, we are using protein interactions that have been identified with high-throughput yeast two-hybrid (Y2H) screens (Ito, et al., 2001) (for details about the data set, see Chapter 2). Protein-protein interactions detected by Y2H suffer from some known disadvantages. For example, it is possible that two proteins, although able to interact, and therefore reported as positives in a Y2H screen, are never in close proximity to each other within the cell (Van Criekinge and Beyaert, 1999). Besides this location constraint, there is also a time constraint, meaning that a pair of proteins that interact in the Y2H experiment may be expressed at different points in the cell cycle, and therefore never interact *in vivo*. The presence of noisy edges makes it difficult to define a quality measure based on pure topology, such as when using a density measure. Therefore, taking into account annotation regarding molecular function of the proteins and their involvement in biological processes or cellular components is likely to increase the reliability of protein-protein interactions, and thereby reduce the number of false positives. We propose the concept of module cohesiveness, based on both topological and semantic information, to describe clusters

of proteins that are not only densely connected but also perform similar functions or participate in the same biological processes. For that purpose, we use the terms stored in biological ontologies. There are different biological aspects that may be covered in ontologies. In this chapter, we describe an integrated measure that is useful for quantifying the topological and semantic cohesiveness. To test the potential of this measure for revealing modular formations, we have applied a module-identifying algorithm (see section 6.2.2). Besides this, we also evaluate which of the two aspects, GO molecular function and GO biological process, is most appropriate for identifying modules in PINs, and if it would be beneficial to combine those aspects.

6.1 Materials and methods

In the following sections we introduce the weighted metrics that arise from the combination of semantic similarity and PIN-topology information and demonstrate their use in analysing the global properties of PINs (see 6.1.1-6.2.1). Furthermore, we describe the algorithm for module identification that is based on one of the matrices, and the results of its application on the yeast PIN based on the CORE data set (see 6.2.2-6.3).

6.1.1 Weighted clustering coefficient

As pointed out in previous work, the individual edge weights do not provide a general picture of the network's complexity (Yook, et al., 2001). Therefore, we here consider the sum of all weights between a particular node and its neighbours, also referred to as the node strength. The strength s_i of the node i is defined as:

$$s_i = \sum_{\forall j,\, j \in N(i)} ss_{ij} \qquad \textbf{(14)}$$

where ss_{ij} is semantic similarity (see Equation 5 on page 35) between nodes i and j, based on their GO terms.

Recently, some extensions of the topological clustering coefficient have emerged for weighted networks. In (Barrat, et al., 2004) two metrics that combine topological and weighted characteristics – weighted clustering coefficient (c^w) and weighted average nearest-neighbours degree (nn^w) – were introduced. These measures have previously been applied to two types of complex weighted networks, namely, the world-wide

airport network and the scientist collaboration network. We introduce a weighted measure that uses semantic similarity weights.

Weighted clustering coefficient c^W is defined as:

$$c_i^W = \frac{1}{s_i(k_i - 1)} \sum_{\forall j.h \mid \{j.h\} \in K(i)} (ss_{ij} + ss_{ih}) \quad (15)$$

where s_i is the functional strength of node i (see Equation 14 on page 80) and ss_{ij} is the semantic similarity reflecting the functional weight of the interaction. For each triangle formed in the neighbourhood of node i, involving nodes j and h, the semantic similarities ss_{ij} and ss_{ih} are calculated. Hence, not only the number of triangles in the neighbourhood of the node i is considered but also the relative functional similarity between the nodes that form those triangles, with regard to the total functional strength of the node. The normalisation factor $s_i(k_i - 1)$ represents the summed weight of all edges connected from node i, multiplied by the maximum possible number of triangles in which each edge may participate. It also ensures that $0 \le c^W \le 1$. It should be noted that we calculate two c^W values for each node, one based on GO molecular function and the second based on GO biological process. We then use the higher of the two as the final weight of the node. This gives the added advantage of taking both aspects into consideration.

To investigate the effect of semantic weights on the overall weighted PIN architecture, we established a weighted analogue to the topological clustering function $C(k)$. The weighted clustering function $C^W(k)$ is defined as the weighted clustering coefficient averaged over all nodes with connectivity k. C^W is defined as the weighted clustering coefficient averaged over all nodes.

By comparing topological and weighted measures, it is possible to provide additional information on the network architecture. In real networks, we often find either $C^W > C$, meaning that the triangles in the network are more likely to be formed by edges with large weights, or $C^W < C$, meaning the opposite (Barrat, et al., 2004).

6.1.2 Weighted nearest-neighbours degree

Other measures used to investigate the topological properties of the network are average degree of the nearest neighbours nn_i, and its weighted counterpart nn_i^w defined as:

$$nn_i^w = \frac{1}{s_i} \sum_{\forall j, j \in N(i)} ss_{ij} k_{ij} \quad (16)$$

In this equation, the weighted average of the topological nearest-neighbours degree is calculated according to the normalised weight of the connecting edges.

To study global properties, we use nn_i and nn_i^w averaged over all nodes with connectivity k. The topological $nn(k)$ may be used to identify two general classes of the networks. If $nn(k)$ is a decreasing function of k, it indicates that nodes with high degree have a higher probability to connect to the nodes with low degree, whereas low-degree nodes tend to connect to the nodes with high degree. This property is referred to as *disassortative mixing*, and it has been observed earlier in cellular networks (Maslov and Sneppen, 2002). In contrast, many networks, for example social networks, show *assortative* behaviour, i.e., a preference for high-degree nodes to connect to other high-degree nodes (Newman, 2002). The advantage of the weighted analogue is that it measures the affinity of the node in terms of its tendency to connect to high-degree or low-degree neighbours, according to the magnitude of the functional similarity of the actual interactions.

The aim of the work described in this chapter is to investigate the weighted measures for describing the network properties, with the purpose of shedding light on how the functional strength of interactions affects our view of the global structural organisation of the protein networks. Besides this aim, we also focus on the main purpose of this work, namely using one of the weighted measures to elucidate densely and functionally connected substructures within the network, which we refer to as modules.

6.2 Results

6.2.1 Structural organisation of PINs enriched with functional weights

A high density of proteins forming triangles is reflected by a high average clustering coefficient C and the presence of this property is a signature of the network's potential modularity (Watts and Strogatz, 1998). For most real networks, C is considerably

higher than that of random networks of the same size (Ravasz and Barabasi, 2003; Watts and Strogatz, 1998). To show how strong the clustering tendency is in the network analysed here, we compared its C with the expected C for a corresponding random network. In previous work (Watts and Strogatz, 1998), it was found that random networks have $C \approx \bar{k}/N$, which here equals 0.002 (since $\bar{k} = 4.8$ and $N = 2231$). We found that $C \approx 0.3$ for the yeast PIN, which is considerably higher than the expected C of a corresponding random network. Furthermore, we compared the clustering function $C(k)$ with the corresponding weighted clustering function $C^{W}(k)$, to gain some additional knowledge about the network's architecture (see Figure 17 on page 84).

To analyse the difference between the distributions of the weighted and topological data sets, we started by performing two nonparametric statistical tests, Kolmogorov-Smirnov and Wilcoxon signed rank test. We compared the cumulative distributions of the weighted and topological data sets using the parameter-free Kolmogorov-Smirnov (http://www.physics.csbsju.edu/stats/KS-test.n.plot_form.html) test to see if the distributions are identical. We could reject the null hypothesis H_0 ($P < 0.001$), meaning that the two distributions differ significantly. To be able to safely conclude that the difference between distributions is significant, we also used the statistical software package R (http://www.r-project.org/) to compute the degree of significance with the Wilcoxon signed rank test. The H_0 that the difference ($d = c^{W} - c$) has a median value of zero or below could be rejected ($P < 2.2 \cdot 10^{-16}$), meaning that we can safely conclude that the weighted clustering coefficient values are higher than the topological clustering coefficient values.

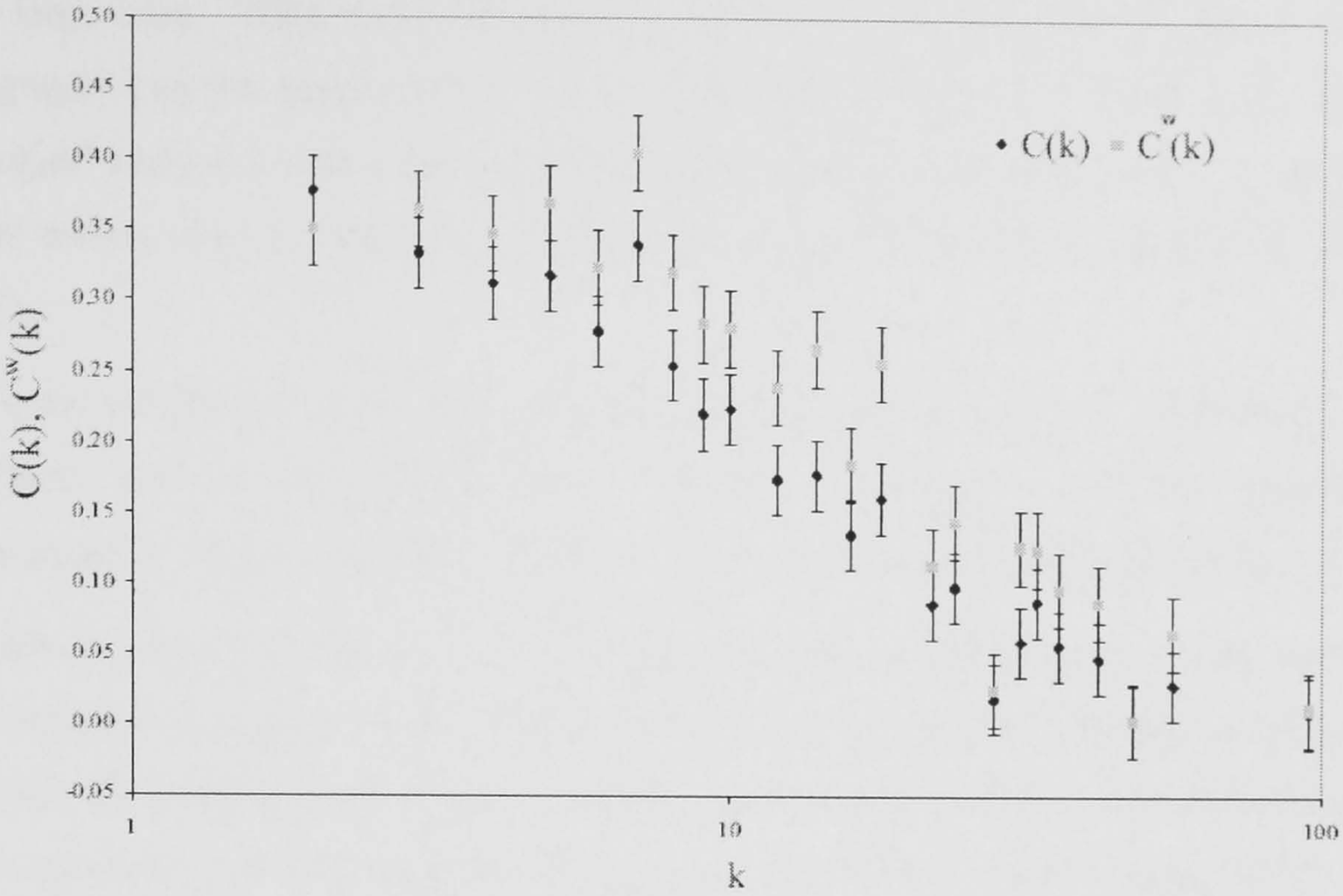


Figure 17: Comparison between topological and weighted clustering function

By analysing the topological clustering function in Figure 17 above, we can notice that the PIN has a decaying $C(k)$, implying that the clustering coefficient is much higher for proteins that have low connectivity, compared with the proteins with high connectivity, i.e., hubs. The role of hubs is to link different, and otherwise not communicating, densely interconnected clusters, i.e., modules. This role accounts for the low clustering coefficient of hubs. The fact that $C^w / C \approx 1.1$, or more specifically, $C^w(k) > C(k)$ for $k > 2$, means that the proteins forming "triangles" are more likely to share high functional similarity with each other. As already stated in (Barabasi and Oltvai, 2004), from a network theory perspective, each module can be reduced to a set of triangles. This statement, together with the obtained results that confirm the accumulation of high functional similarity in highly interconnected protein sub-graphs, supports the hypothesis that we are dealing with a highly modular network.

The comparison between the weighted measure and its topological analogue shows that the weighted clustering coefficient values are generally higher than the values of their topological counterpart (see Figure 17). The overlap between the standard error bars in Figure 17 indicates that the difference between the weighted and topological values is not statistically significant for $k = 3$ and $k = 4$ (but that is significant for $4 < k < 22$). This may be attributable to the nature of the network or the experiments used to generate it. As mentioned before, small cliques are more likely to emerge by chance

than large ones. Thus, the functional relevance of the smallest highly interconnected sub-graphs may be questioned. This may also reflect the theory about modularity in biological networks which states that the most relevant functional activities correspond to the meso-scale (5-25 proteins) rather than to the entire network (Spirin and Mirny, 2003).

The other weighted measure that we used to investigate the network's architecture, i.e., weighted average nearest-neighbours degree, is compared with its topological counterpart in Figure 18 below. We can observe that the topological $nn(k)$ has an overall decreasing tendency, thus reflecting the disassortative nature of the network. This indicates that nodes with large connectivity (hubs) "avoid" linking to each other directly, and instead connect to the proteins that have few partners. This behaviour has been reported for protein networks in previous work (Maslov and Sneppen, 2002). The comparison between the weighted and topological function in Figure 18 reveals that $nn^{w}(k) > nn(k)$ for all degrees, with the exception of $k = 2$. This reflects the tendency of edges with high semantic weights to connect with neighbours with high degrees.

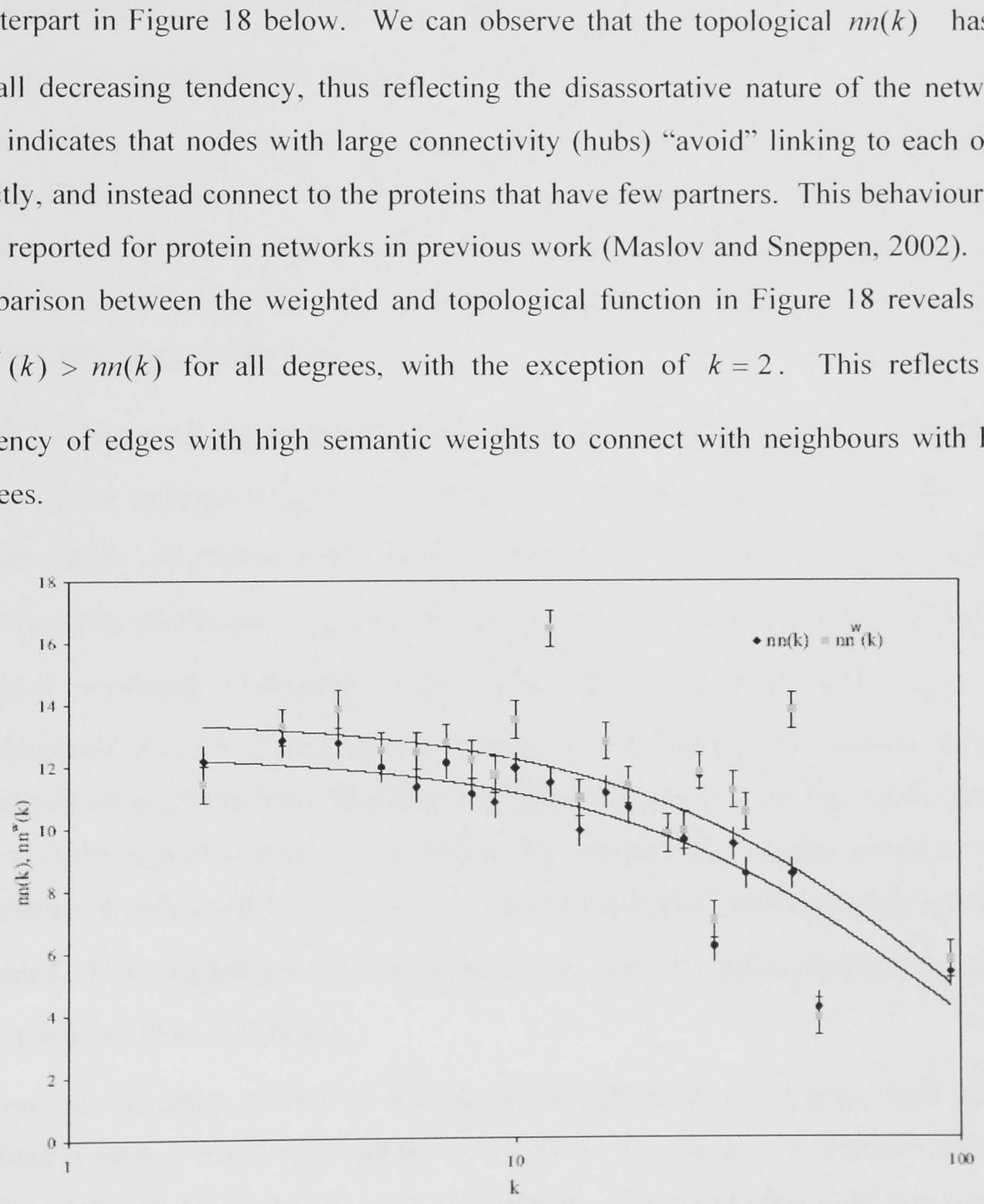


Figure 18: Topological and weighted average nearest-neighbours degree

6.2.2 The algorithm for module identification

The aim of the study described in this chapter is to identify highly interconnected sub-graphs with high functional homogeneity. We call those sub-graph modules. In previous work by Bader and Hogue (2003), an algorithm for finding complexes in large-scale networks, called MCODE, has been developed. MCODE is based on the weighting of nodes with a so called core-clustering coefficient. The core-clustering coefficient of a node i is defined as the density of the highest k-core in the closed neighbourhood of i, i.e., $N[i]$. A k-core of a graph is a sub-graph containing a set of nodes, each of which is connected to at least k other nodes in the set.

In this work, we propose an alternative algorithm, called SWEMODE (Semantic WEeights for MODule Elucidation), that is used for deriving functional modules, based on the functional and topological cohesiveness of the sub-graphs. The first stage of the algorithm is node weighting. We define two weighting schemes that are based on weighted clustering coefficients.

The first scheme uses the weighted max-clustering coefficient ($c^{w}_{\max}$), which we define as the highest weighted clustering coefficient in the closed neighbourhood $N[i]$. The relative weight assigned to node i based on this value, is the product of the weighted max-clustering coefficient $c^{w}_{\max}$ and the connectivity of the node that has the highest weighted clustering coefficient. This connectivity is denoted with $n_{\max}$ (see pseudocode in Text box 2 on page 88). If there is a tie between two or more nodes in the immediate neighbourhood regarding the highest weighted clustering coefficient, the node with the highest connectivity is chosen. By assigning this relative weight to i, the importance of nodes with high degree, participating in highly interconnected regions, is amplified. This weighting scheme is referred to as $\max(c^{w})$ and is used as an example in the pseudocode in Text box 2.

The second weighting scheme is used to test the effect of multiplying each node's weighted clustering coefficient with the connectivity of the node. The purpose of this is to enhance the weight of densely connected proteins. We refer to this weighting scheme as $\mathrm{dens}(c^{w})$. Both weighting functions were chosen because they combine functional and topological information, which we consider to be an advantage. In addition, we have also derived modules using a purely topological weighting scheme based on the

topological clustering coefficient c. The relative weight, assigned to node i, is the product of the highest clustering coefficient $c_{\max}$ in $N[i]$ and the connectivity of the node that has the highest c. This weighting scheme is referred to as $\max(c)$. There are other functions, such as the density function (Bader and Hogue, 2003), but these are not evaluated here.

The second stage of the algorithm, i.e., core module prediction, is similar to the molecular complex prediction step of MCODE (Bader and Hogue, 2003). It uses the node weights, seeds a module with the highest weighted node, and then moves through the neighbourhood of the seed node, identifying neighbours in the immediate neighbourhood whose weights satisfy the node weight percentage (*NWP*) requirement, and including those nodes in the module. This module prediction procedure is repeated using the next available node (a node that has not already been added to a module) with the second highest weight as a seed for a new module, and so on until the end of the node ranking. The requirement for inclusion of the neighbours in a module is that their weights are higher than a threshold, which is a given *NWP* of the seed node (Bader and Hogue, 2003).

SWEMODE has three options concerning traversal of nodes that are considered for inclusion in a module. The first option, immediate neighbour search, only considers the immediate neighbours of the seed node. The second option is to traverse the protein graph starting from the seed node, using depth-first search (DFS), and to add nodes to a module according to the given criterion regardless of their distance from the seed node. The third option is a depth-limited search, where the search is limited to a certain distance from the seed node, by setting the desired maximum depth of the neighbourhood in which the search is allowed. In the initial experiments described in this chapter, we focused in our analysis on the modules obtained from immediate neighbours search only, because those modules showed the highest agreement with the MIPS complexes used in the evaluation (see section 6.2.3). This is not surprising because we used complexes at the lowest level of the hierarchy. For example, the 26S proteasome complex class contains 36 proteins, and consists of two sub-classes at the lowest level in the hierarchy, 20S proteasome (15 proteins) and 19/22S regulator (18 proteins), which we use in the evaluation. When including indirect neighbours in module prediction (by using DFS), this generally results in larger modules that stretch over several complexes, and in this case shows higher agreement with the complex at a

higher level in the hierarchy (26S proteasome). The later option is also important to consider when we analyse the potential interconnectivity between modules.

Step 1: Node Weighting

Procedure *Node_Weighting*

▷ Input: $G = (V, E)$

▷ $N[i]$: closed neighbourhood of i

▷ $c^{w}_{\max}$: highest weighted clustering coefficient in $N[i]$

▷ $n_{\max}$: connectivity of the node in $N[i]$ with $c^{w}_{\max}$

For each $i \in G$

Identify $N[i]$

Get $c^{w}_{\max}$

Get $n_{\max}$

Set weight of $i = c^{w}_{\max} \cdot n_{\max}$

end for

end procedure

Step 2: *Module_prediction*

Procedure *Get_Module*

▷ Input: $G = (V, E)$; node weights W; node weight percentage *NWP*; seed node j:

▷ M : module vector

If j already visited **then return**

else

for all $i \in N(j)$

if (weight of i > (weight of $j \cdot NWP$)) then add i to M

end for

end procedure

Text box 2: Algorithm for module identification

6.2.3 Evaluation of SWEMODE using MIPS complexes

SWEMODE was used to predict functional modules in the CORE data set. Resulting modules were then compared with the MIPS data set of known protein complexes. The MIPS (http://www.mips.gsf.de/proj/yeast/) protein complex catalogue is a curated set of manually annotated yeast protein complexes derived from literature scanning. After removal of 44 complexes that contain only one member, 212 complexes were left in the data set.

It should be pointed out that the MIPS complex data set is incomplete, which may have affected the presented outcome in terms of the number of matched complexes. For example, the complex containing Lsm-proteins, which has the highest ranking in our evaluation (see section 6.2.5), is not present in the MIPS complex data set, although it is a well-known complex (Fromont-Racine, et al., 2000; He and Parker, 2000; Rader and Guthrie, 2002). Furthermore, a module may consist of a protein complex and some additional proteins that interact with the complex to perform a distinct function.

Even though the MIPS complex data set is incomplete, it is currently the best available resource for protein complexes that we are aware of. We are convinced that future applications of this work will contribute to developing a benchmark that can be used for a more thorough evaluation of prediction accuracy.

In the following experiments, SWEMODE was run using three different weighting schemes. Two of the weighting schemes, $\max(c^W)$ and $\text{dens}(c^W)$, were based on weighted clustering coefficient, whereas the third one, $\max(c)$, was based on the topological clustering coefficient. As we explained earlier in the section 6.1.6, we combine two GO aspects by calculating two weights for each node, one based on GO molecular function, and the second based on GO biological process. We then use the higher of the two as the final weight of the node. This gives the added advantage of taking both aspects into consideration.

Concerning traversal of the nodes in the module prediction step, SWEMODE was run with all three options (immediate neighbours search, DFS, and depth-limited search) over a range of 20 *NWP* parameter values (0 to 0.95 in increments of 0.05). The neighbourhood depth parameter used in the depth-limited search, was varied between 1 and 3. The best results for all three weighting schemes were obtained when the depth

parameter was set to 1, i.e., when only immediate neighbours were considered for inclusion in a module.

To evaluate the performance of SWEMODE and choose the best parameter settings, we used two different scores: overlap score and density score. In previous work, a similar evaluation has been applied to the clustering algorithm MCODE (Bader and Hogue, 2003), with respect to the number of matched complexes, but we here use another definition of overlap score (see Equation 17 below). The best choice of parameters for SWEMODE is the one that predicts the largest number of modules that match MIPS protein complexes (at the high overlap score threshold levels), and the most densely connected modules.

The overlap score Ol (Poyatos and Hurst, 2004), which has been defined earlier for the purpose of finding cluster overlap (see Section 5.1.3), is here reused:

$$Ol_{ij} = \left|M_i \cap M_j\right| \Big/ \sqrt{\left|M_i\right|\left|M_j\right|} \tag{17}$$

where M_i is the predicted module, and M_j is a module from the MIPS complex data set. The Ol measure assigns a score of 0 to modules that have no intersection with any known complex, whereas modules that exactly match a known complex get the score 1. We use the overlap score to decide upon the choice of weighting schema. The result for all three weighting functions with respect to the maximum number of matched complexes is presented in Figure 19 on page 91. In Figure 19, the maximum number of matched complexes for three weighting schemes is plotted as a function of overlap score threshold. The maximum number of matches at each overlap threshold level Ol is based on module sets derived over the range of 20 NWP parameter values. As Ol increases, fewer predicted complexes match known complexes. By using dens(c^W) and only considering immediate neighbours for inclusion in modules, we obtained the results that we consider to be the most promising.

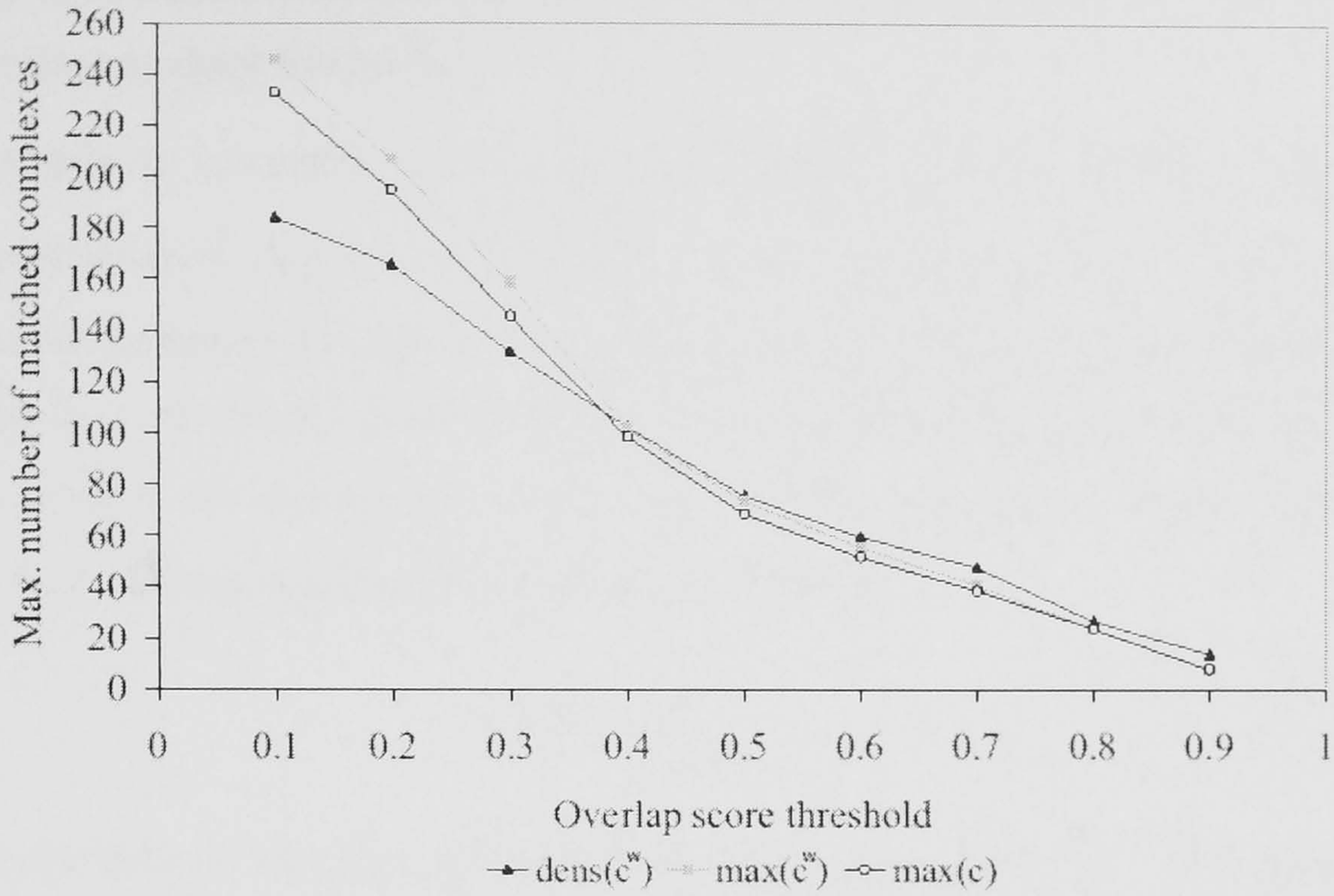


Figure 19: Evaluation of three weighting schemes with overlap score threshold

Because SWEMODE predicts fewer modules when $\text{dens}(c^W)$ is used, this results in fewer modules that pass the threshold of 0.1, compared with the other two weighting schemes. It should be noted that 93 of the modules on average, predicted with $\max(c^W)$, and 94 modules on average, predicted with $\max(c)$, did not pass the threshold level $Ol > 0$, meaning that they have no similarity with any of known MIPS complexes. For $\text{dens}(c^W)$, the corresponding number was 66. This may indicate that the two other weighting schemes introduce larger numbers of false positives.

The largest number of modules that pass the threshold of 0.1 is generated when using the other weighted scheme $\max(c^W)$. However, $\text{dens}(c^W)$ results in a larger number of matched complexes, compared to the two other schemes, when the overlap score threshold is considerably high ($Ol > 0.4$). For example, with $\text{dens}(c^W)$, 14 modules are identified that perfectly match MIPS protein complexes ($Ol > 0.9$). The corresponding number for both $\max(c^W)$ and $\max(c)$ is 8. Further analysis was focused on the two weighted schemes, because those performed better than the topological one.

As mentioned earlier, protein modules are characterised by the property that their members have high rates of interaction with each other. We therefore consider module

density to be a useful criterion when deciding which weighting scheme performed better in revealing modular formations.

The density is calculated for the modules generated with SWEMODE using two weighting schemes $\text{dens}(c^w)$ and $\max(c^w)$. Given a module graph $G = (V, E)$, where the number of nodes (proteins) is denoted by k, and the number of edges (interactions) is denoted by n, the density is defined as n divided by the theoretical maximum number of edges possible for the module graph, $n_{\max}$ (Bader and Hogue, 2003), defined as $k(k-1)/2$. Hence, the density for a module is defined as:

$$Density = \frac{n}{n_{\max}} \tag{18}$$

The result from the density comparison is shown in Figure 20 below. The average and maximum density values are plotted as function of the *NWP* parameter values. The modules predicted with $\text{dens}(c^w)$ have higher density than the modules predicted with $\max(c^w)$, for $0 < NWP < 0.85$, which also supports our decision to choose the modules predicted with $\text{dens}(c^w)$ for further analysis. The density for modules predicted with $\text{dens}(c^w)$ starts to decrease for $NWP > 0.6$.

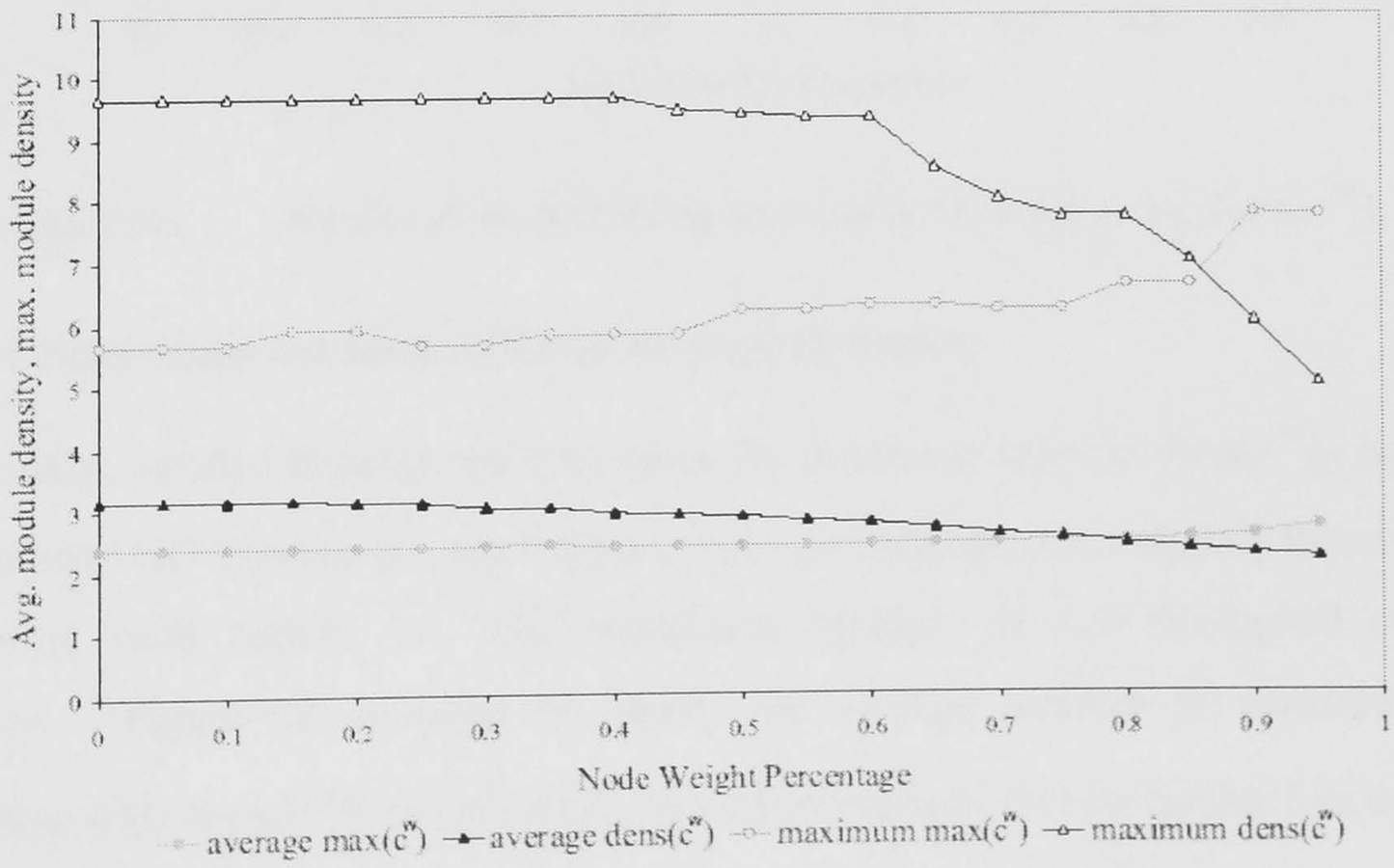


Figure 20: Average and maximal module density

The next task in the analysis was to find the best module set, generated with dens(c^W), as this weighting scheme was identified as the best performing one. The best results in terms of number of matched complexes is obtained for $NWP > 0.4$. Figure 21 below shows the number of modules that matched complexes through all NWP values, at the overlap score threshold level $Ol > 0.4$. The highest number of predicted modules that match MIPS modules, 103, was obtained with $NWP > 0.4$. By comparing Figure 21 with Figure 20, that shows the average and maximum density values for the same range of NWP values, we can observe that a decrease of the number of modules that match MIPS complexes implies a decrease in module density.

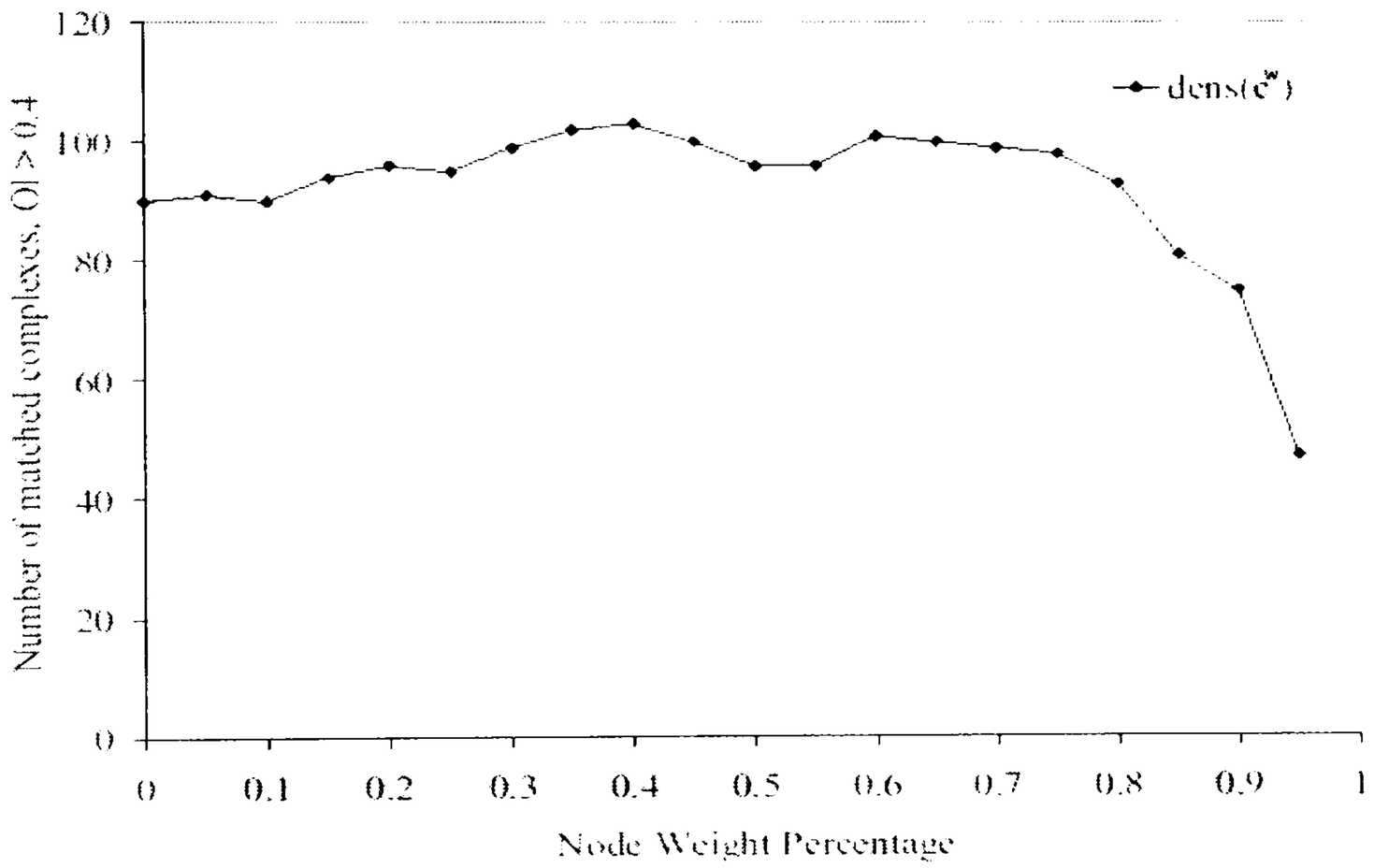


Figure 21: Number of matched MIPS complexes at $Ol > 0.4$ using dens(c^W)

6.2.4 Comparison between different biological aspects

In this study, we also demonstrate that using the weighting scheme dens(c^W), based on the combined GO aspects, i.e., the higher of the two weights, gives slightly better results than using each aspect, i.e., GO molecular function or GO biological process, separately. Figure 22 on page 94 shows the average number of matched MIPS complexes with dens(c^W) based on GO biological process, GO molecular function, and combination of both. The number of matched complexes is calculated over the range of 20 NWP values and plotted as a function of overlap score threshold.

We also calculated the average overlap between all sets of process-based modules generated over a range of 20 *NWP* values with the corresponding sets of function-based modules. The average overlap O, is defined as:

$$O = \frac{1}{|M_p|} \sum_{i=1}^{M_p} \max\left\{ Ol_{i,j \; j=1..|M_f|} \right\} \quad (19)$$

where $|M_p|$ is the number of modules based on biological process annotation, $|M_f|$ is the number of modules based on molecular function annotation, and Ol is the overlap score as defined in Equation 17 on page 90. The highest average overlap $O = 0.76$ is obtained with $NWP > 0.4$. As expected, there is a significant overlap between modules generated with the two different types of ontology, and the same parameter setting that resulted in the highest average overlap also generates the best results with respect to the number of matched MIPS modules.

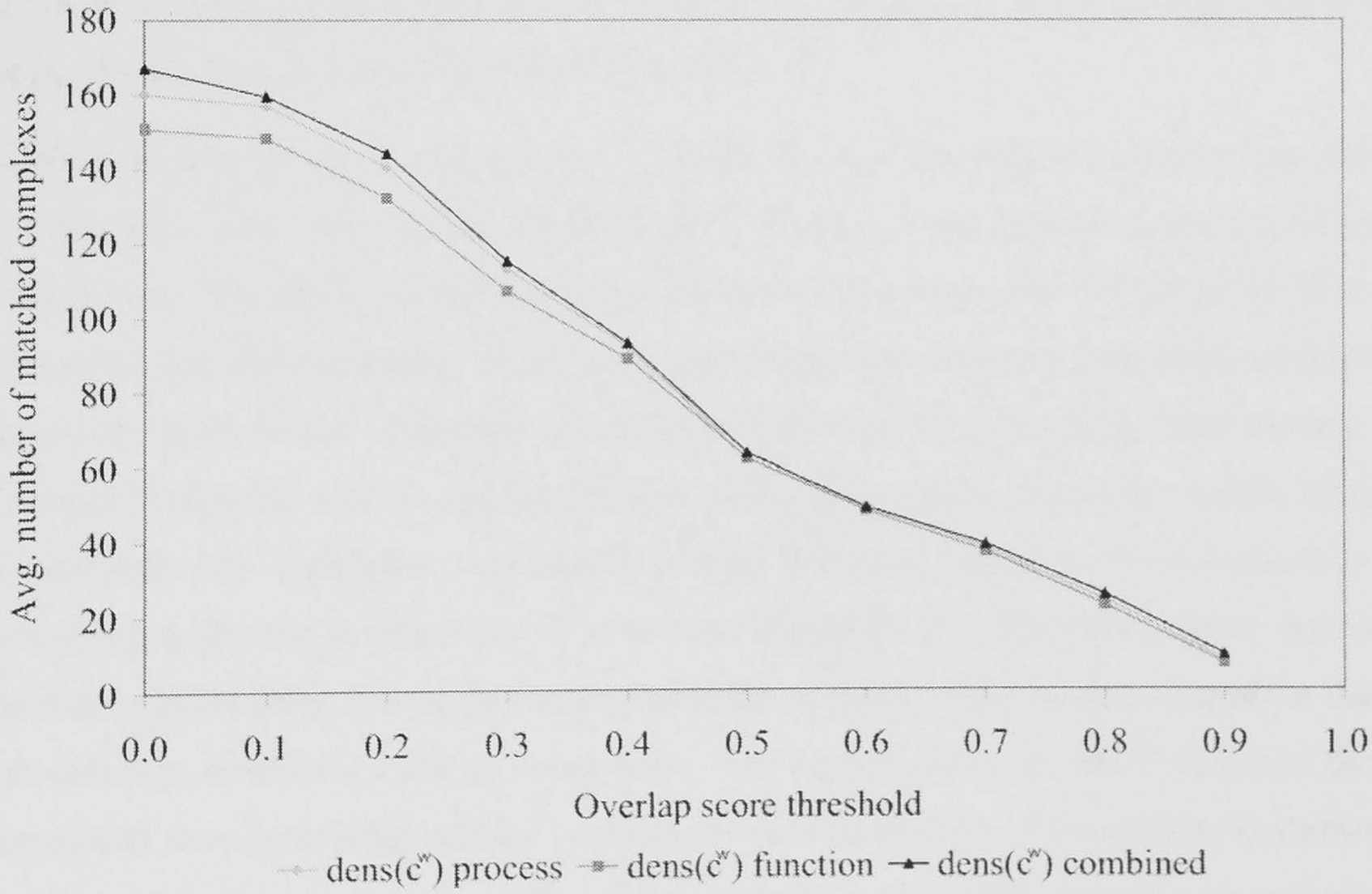


Figure 22: $\text{dens}(c^w)$ based on combined GO aspects compared to each aspect

6.2.5 Analysis of potential modules

SWEMODE was applied on 2231 proteins from the CORE yeast data set, with 6375 interactions (see section 2.7.2). Different values of the *NWP* threshold parameter were

tested, using three different node weighting schemes (see section 6.2.2). The results for the weighting scheme dens(c^W) when $NWP > 0.4$ were selected as the most promising, because this combination predicted the largest number of modules with considerably high overlap with MIPS complexes. The chosen *NWP* criterion means that the weight of a neighbour must be larger than 40% of the weight of the seed node. Lower values of this parameter resulted in inclusion of unknown proteins or proteins that are only weakly functionally related to the seed protein. However, higher values of this parameter resulted in many modules with few proteins, with the majority consisting of only two or three proteins. When using 0.4 as threshold, 236 modules were identified (modules containing only one member were not considered). Because the analysis of the global properties reveals that the nodes with the smallest connectivity have weaker functional weights on the edges that form triangles, and because cliques of size three or less may be more likely to emerge by chance (Spirin and Mirny, 2003), we have excluded those modules from further analysis. Thereafter, a total of 81 modules were left in the analysis. The complete list of these 81 modules can be found in Table 13 (see Appendix B), sorted in the order of decreasing density.

Table 7 on page 98 shows a list of the 15 top ranked modules and, for comparison, the five modules with lowest rank. Module rank is based on the density score, shown in column two. The density of the module is multiplied with the number of its members to obtain the final density score. In this way, larger and more densely connected modules are given higher scores. The third column shows the number of proteins in the module. Common functional activity for the proteins within the module, shown in column four, is represented by their most significantly shared GO term, based on the sub-ontology describing molecular function (for *P* values see Appendix B). The fifth column shows the most significantly shared GO term among the proteins in the module, based on the sub-ontology describing cellular component. The significance (i.e., the *P* value) of the shared GO term describing cellular component for the members of the module is shown in column six. SGD GO Term Finder (http://db.yeastgenome.org/cgi-bin/GO/goTermFinder) was used to calculate the *P* values. Column seven, Frequency, shows the percentage of the proteins within the module that are annotated with the given GO term.

Module rank	Score	Proteins	Functional activity	Cellular component	*P* value	Frequency (%)
1	9.67	13	RNA binding	Small nuclear ribonucleo-protein complex	$5.72 \cdot 10^{-21}$	90
			Lsm5, Lsm6, Lsm2, Lsm4, Lsm1, Lsm8, Lsm3, Lsm7, Pat1, Prp4, Prp24, Smd2, Smd3			
2	7.88	18	Endopeptidase activity	Proteasome complex	$1.10 \cdot 10^{-30}$	83
			Rpt1, Rpt2, Rpt3, Rpt4, Rpt5, Rpt6, Rpn1, Rpn3, Rpn10, Rpn11, Rpn12, Rpn14, Cdc6, Pre1, Nas6, Leo1, Ctr9, Rad23			
3	7.71	8	Dolichyl-diphosp-hooligosaccharide-protein glycotra-nsferase activity	Oligosaccharyl transferase complex	$5.39 \cdot 10^{-21}$	100
			Wbp1, Ost1, Ost2, Ost3, Ost4, Ost5, Swp1, Stt3			
4	7.64	12	RNA binding	mRNA cleavage factor complex	$1.81 \cdot 10^{-31}$	100
			Cft2, Pap1, Pfs2, Fip1, Pta1, Pti1, Mpe1, Pcf11, Ref2, Rna14, Swd2, Ssu72			
5	7.54	14	3'-5'-exoribonuclease activity	Transcription factor complex	$1.34 \cdot 10^{-25}$	100
			Cdc39, Cdc36, Not3, Not5, Mot2, Caf4, Ssn3, Ssn2, Pop2, Ccr4, Caf40, Caf130, Taf1, Taf6			
6	7.00	13	ATP-dependent RNA helicase activity	Nucleolus	$2.83 \cdot 10^{-16}$	92
			Nop15, Erb1, Mak21, Brx1, Dbp10, Has1, Nop4, Nop7, Rlp7, Sda1, Nug1, Cic1, Ytm1			
7	6.86	8	Structural constituent of cytoskeleton	Septin ring	$2.11 \cdot 10^{-15}$	75
			Cdc11, Cdc3, Cdc10, Cdc12, Shs1, Kcc5, Gin4, Bni5			
8	6.86	8	NAD-independent histone deacetylase	Histone deacetylase complex	$9.32 \cdot 10^{-18}$	88

			activity			
		Hos2. Snt1. Hos4. Set3. Hst1. Zds1. Sif2. Cpr1				
9	6.57	8	Translation initiation factor activity	Eukaryotic translation initiation factor 2B complex	$8.47 \cdot 10^{-15}$	62
		Gcd1. Gcn3. Gcd2. Gcd6. Gcd7. Gcd11. Sui2. Sui3				
10	6.44	10	Structural molecule activity	Pore complex	$5.12 \cdot 10^{-17}$	89
		Nup84. Nup145. Nup85. Nup42. Nup100. Nup120. Msn5. Seh1. Sec13. Nup57				
11	6.33	7	DNA clamp loader activity	DNA replication factor C complex	$9.13 \cdot 10^{-21}$	100
		Rfc3. Rfc4. Rfc2. Rfc5. Cft8. Cft18. Elg1				
12	6.00	7	Hydrogen-transporting ATP synthase activity. rotational mechanism	Proton-transporting ATP synthase complex	$1.16 \cdot 10^{-18}$	100
		Atp18. Atp6. Atp1. Atp2. Atp7. Atp17. Atp20				
13	6.00	6	DNA replication origin binding	Origin recognition complex	$3.10 \cdot 10^{-19}$	100
		Orc6. Orc5. Orc1. Orc2. Orc3. Orc4				
14	6.00	6	Ubiquitin-protein ligase activity	Anaphase-promoting complex	$1.11 \cdot 10^{-16}$	100
		Apc1. Doc1. Cdc16. Cdc23. Cdc26. Cdc27				
15	5.73	16	snoRNA binding	Small nucleolar ribonucleoprotein complex	$1.13 \cdot 10^{-22}$	75
		Utp22. Utp4. Utp6. Utp8. Utp7. Utp10. Utp18. Utp21. Pwp2. Prp45. Dip2. Ecm16. Emg1. Kre33. Rok1. Enp2				
77	2.00	4	GTPase activity	No significant ontology term	-	-
		Sar1. Mf(alpha)1. Cdc7. Tem1				

78	2.00	6	No significant ontology term	Nucleolus	$6.72 \cdot 10^{-07}$	83
		Nsa2. Nog1. Nog2. Noc2. Ycro723. Mrt4				
79	2.00	4	Small GTPase regulator activity	Actin cap	$2.53 \cdot 10^{-05}$	50
		Cla4. Gic1. Rga1. Boi2				
80	2.00	4	Binding	Cystol	0.038	50
		Kap123. Kap104. Sec7. Yap1				
81	2.00	4	No significant ontology term	Golgi transport complex	$1.44 \cdot 10^{-12}$	100
		Cog2. Cog1. Cog6. Cog4				

Table 7: Statistics for top 15 modules and bottom 5 modules. For explanations, see text

The most significantly shared term, as determined by GO Term Finder, is obtained by examining the group of proteins to find the GO term to which the highest fraction of the proteins is associated compared to the number of times that the term is associated with other yeast proteins. In addition, every second row of the table lists the names of the proteins included in the corresponding module.

The functional module with highest rank corresponds to the Lsm complex. Figure 23 on page 99 shows all member proteins of this module and their interactions. When generating this module, Lsm5 was used as seed protein (all proteins that have been used to seed modules are underlined in the table). All eight Lsm-proteins are correctly predicted by the algorithm. Sm-like (Lsm) proteins participate in a variety of RNA processing events. For example, Lsm1-Lsm7 are involved in mRNA degradation and splicing (He and Parker, 2000). Besides Lsm-proteins, this functional module contains Prp4 and Prp24, which are splicing factors with functionally significant interactions with Lsm-proteins (Fromont-Racine, et al., 2000). These interactions are necessary for pre-mRNA splicing (Rader and Guthrie, 2002). Three remaining proteins in the module (Smd2, Smd3 and Pat1) are also closely related to the other proteins in the module. Smd2 and Smd3 are, along with the Lsm-proteins, part of the small U4/U6.U5 nuclear ribonucleoprotein complex (Stevens and Abelson, 1999). In addition, Lsm1-Lsm7

interact with Pat1, which is a decapping activator. In this way, the Lsm-proteins may promote mRNA decapping, which is necessary for mRNA degradation (Tharun, et al., 2000). However, even if it is apparent that Pat1 has an important role in mRNA degradation, and should be included in this module, it is important to note that this protein is annotated with the term "molecular function unknown" and therefore could not be identified as part of the module when only GO molecular function was considered. In other words, the fact that the SWEMODE algorithm also uses the GO biological process aspect made it possible to place this protein in the correct module. This result suggests the involvement of Pat1 in mRNA degradation.

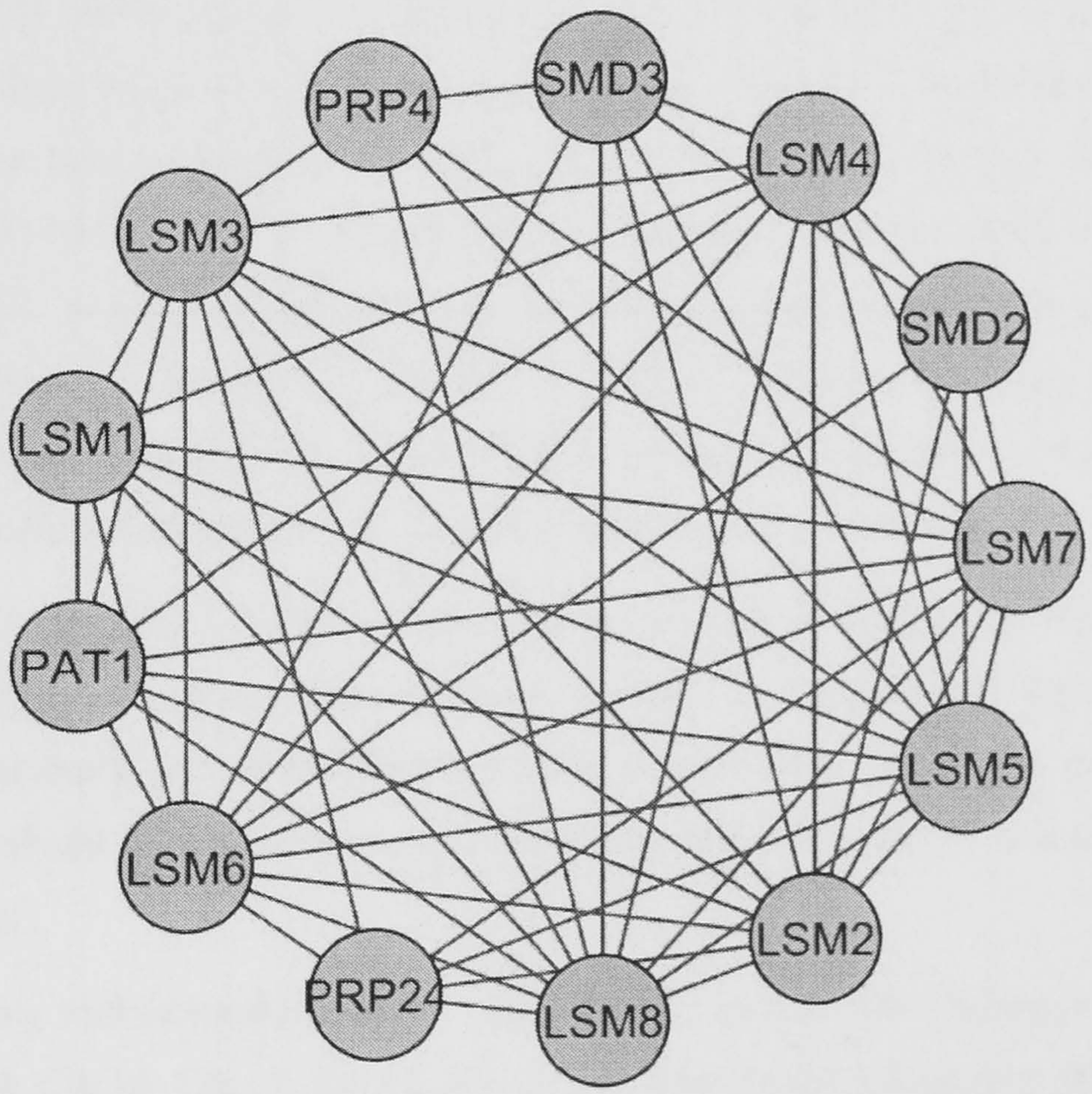


Figure 23: Protein interaction sub-graph containing Lsm proteins. The graph was generated with Cytoscape (www.cytoscape.org)

In previous work by Bader and Hogue (2003), the predicted Lsm-complex lacks Prp24 and Prp4, as expected, because those two proteins are not part of the complex, which implies that they are not so densely connected to other proteins. Therefore, using a graph density function, i.e. a topological function, has proven useful for identifying complexes, but not functional modules. Identifying functional modules, i.e. protein

complexes together with their affectors and regulators, which together form a distinct functional unit, requires inclusion of functional weights. This is one of the examples that demonstrate the distinction between complexes and modules, as well as the advantage of using domain specific knowledge. Because Prp24 is highly functionally related to the Lsm-complex and is required for pre-mRNA splicing to function normally, it forms a functionally distinct module together with the Lsm-complex. Consequently, it was correctly identified by SWEMODE as part of the predicted module.

The module with rank 2 corresponds to the regulatory particle (RP) of 20S proteasome containing two sub-complexes. One sub-complex (the base) contains six ATPases (Rpt1-Rpt6) that are involved in unfolding and translocation of substrates to the 20S proteasome's catalytic chamber (Takeuchi and Tamura, 2004), and Rpn1, which is one of the largest subunits of the proteasome. Furthermore, this functional module also contains Rpn3, as well as Rpn11 and Rpn12, which are parts of the other sub-complex (the lid). Rpn10 is thought to provide a "bridge" between the base and the lid (Wollenberg and Swaffield, 2001) and it is also included in the module. Proteins Rpn5-Rpn9, which also constitute the lid, are not present in the CORE data set.

The module with rank 3 consists of 8 different subunits that form the oligosaccharyl transferase (OST) complex. This complex catalyses a vital step of *N*-glycosylation, which is an essential protein modification (Knauer and Lehle, 1999). The most recently discovered subunit of this complex, Lsm6 (Knauer and Lehle, 1999), was not found by the algorithm.

Another highly ranked module (rank 7) includes five proteins that belong to the Septin family (Cdc3, Cdc10, Cdc11, Cdc12, Shs1), and three septin-interacting proteins Bni5, Gin4, and Kcc4, which are potential septin regulators (Versele and Thorner, 2004). Septin proteins are necessary for proper morphogenesis and cytokinesis (Versele and Thorner, 2004). Also in this example, using only GO molecular function was not sufficient to identify the members of the complex, but this was again compensated by using the higher of the function and process weights (in this case, as in the previously mentioned one, the process weight was the highest of the two).

Figure 24 on page 101 shows the original sub-graph from CORE, containing members of the septin complex and interactions among them, where the width of each edge being proportional to the functional semantic similarity between the proteins that it connects.

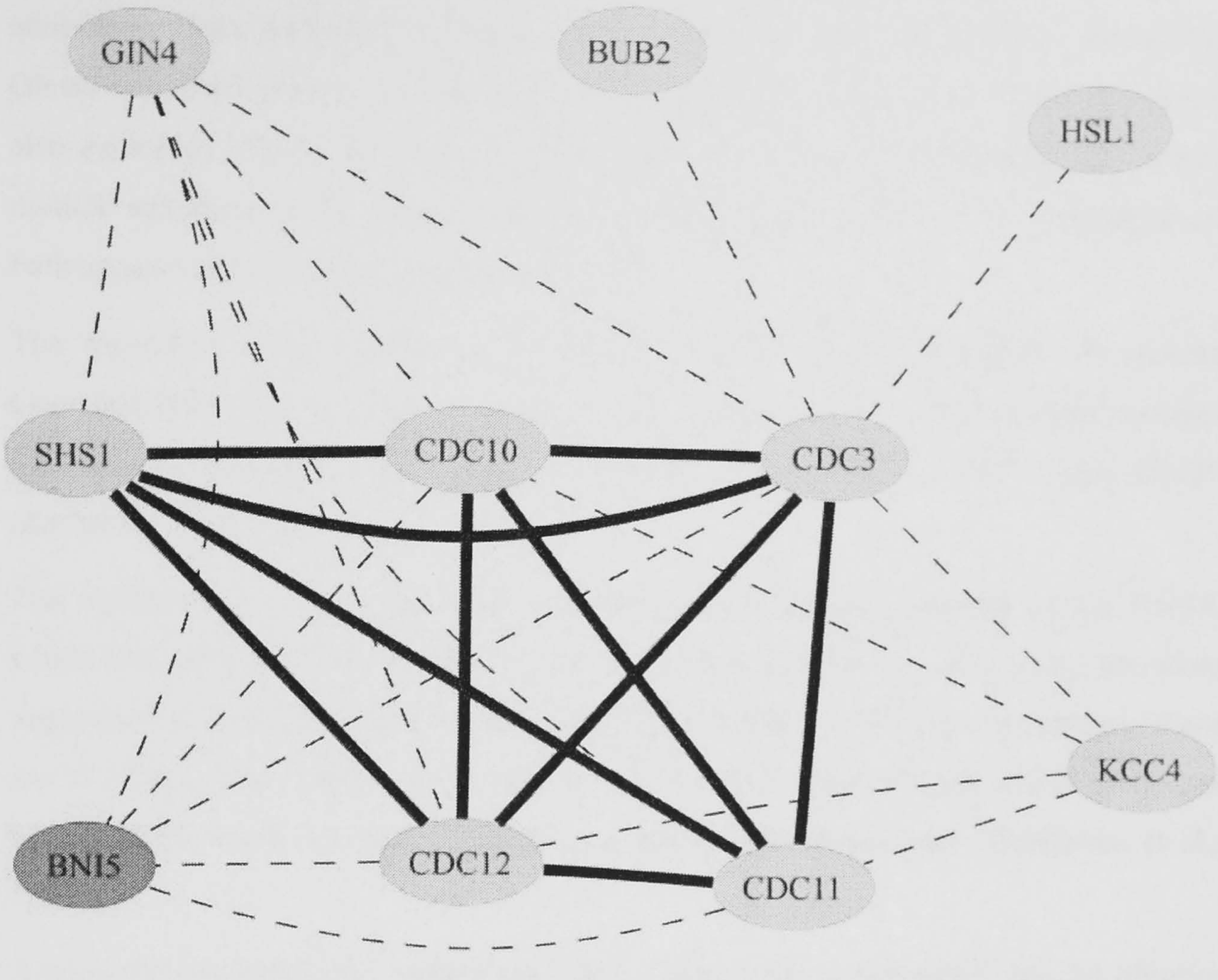


Figure 24: Original protein interaction sub-graph containing septin ring proteins. The graph was generated with GraphViz (www.graphviz.org)

Semantic similarity values were here used as input to GraphViz (http://www.graphviz.org/). Nodes are coloured green if they are annotated with at least one GO function, otherwise red. The width of each edge is proportional to the semantic similarity between the nodes that it connects. Zero-width lines are replaced by dashed lines. Bni5 has unknown function and it does not have any functional similarity with the rest of the module. By using GO functional annotation, only five proteins were identified as members of the module (see the nodes connected with thick lines). Bni5 (red-coloured node) is assigned the GO term "molecular function unknown", since its function is not determined in detail, which is why it would not be identified as part of the module if only GO molecular function would be considered. However, the literature suggests that Bni5-septin interaction is important for septin ring stability and function, which, in turn, is critical for normal cytokinesis (Lee, et al., 2002). This is also confirmed by using GO biological process annotation, which therefore appears to be well-suited for finding modular formations. By including the GO biological process

annotation, three additional members were assigned to this module (Bni5, Kcc4, and Gin4), which are closely associated with septin proteins. However, because there are also examples where GO molecular function facilitates the identification of correct module members, in the cases where GO biological process fails, the combination of both appears to be the best alternative overall.

The algorithm also identified the functional module of the Anaphase Promoting Complex with rank 14, where Apc1 encodes the largest subunit. This complex contains the tetratricopeptide repeat (TPR) subunits Cdc16, Cdc23, Cdc27, and Cdc26 (Zachariae, et al., 1996).

The module with rank 15 contains 16 proteins. All included proteins except Kre33, which has unknown function, are involved in rRNA processing. This result therefore suggests that Kre33 is also involved in rRNA processing. Of those 16 proteins, there are 12 (Pwp2, Utp4-Utp8, Utp10, Utp18, Utp21, Utp22, Dip2, Emg1, and Ecm16) that belong to the small nucleolar ribonucleoprotein (snoRNP) complex (Bernstein, et al., 2004).

Among the five functional modules in Table 7 having the lowest ranks, we can observe that some do not represent statistically significant clusters w.r.t. GO annotation. The cluster frequency for low-ranked modules is generally lower compared with the corresponding value for modules with higher ranks. In the module with rank 80, for example, for three of the four proteins (Kap123, Kap104, and Yap1), the parent GO term "binding" that describes molecular function, has low specificity, resulting in a rather high P value for this module (see Appendix B), compared with the other modules. Only 50% of the proteins in the module share the parent term "cystol", that describes cellular component, and has rather low specificity ($P = 0.038$). Hence, this module is not considered as functionally homogeneous, which explains its low rank.

6.3 Summary and conclusions

In this chapter, we described a method for analysis of protein networks using a weighted clustering coefficient. Weighted clustering coefficient is a novel method that combines functional and topological knowledge, to identify modular formations in protein interaction networks. This measure takes advantage of the semantic similarity between proteins based on Gene Ontology terms, to quantify the functional homogeneity of potential modules in the network. We developed the module-identification algorithm

SWEMODE that incorporates this measure. Many densely connected and functionally cohesive regions have been identified by applying the algorithm. In many cases those regions correspond to sets of proteins that constitute known molecular complexes and some additional interacting proteins which share high functional similarity with the complex but are not part of it. Together, such sets of interacting proteins form functional modules that control or perform particular cellular functions, without necessarily forming a macromolecular complex. Many of the identified modules correspond to the functional subunits of known complexes.

We have also demonstrated that the use of a weighting scheme based on combined GO aspects, i.e., GO molecular function and GO biological process, yields better results than using each aspect separately.

The developed method, SWEMODE, can be used to derive hypothetical functions of an unknown gene product that has physical interactions within a functional module. As indicated by the results, the use of a functionally informed measure to generate modules should imply increased confidence in the predicted function.

There is no doubt that semantic similarity using the GO annotation of proteins is useful in the assessment of the functional similarity between proteins. GO is rapidly becoming the *de facto* standard for gene product annotation in molecular biology. However, it does have some disadvantages that should be mentioned. The depth (specificity) of the GO graph varies for different terms. A reason for weak functional similarity between some interacting proteins may be that their detailed functions have not been determined experimentally. A future more fine-grained GO annotation may lead to improved specificity, i.e. those interacting proteins may be identified as parts of modules. However, it we should not neglect the fact that the drawback of current versions of GO can cause false negatives in our module prediction method.

In Chapter 7, we continue our work by investigating other weighting functions, based on new topological properties and by introducing the GO cellular component annotation. We will also modify the algorithm so that it will allow overlapping modules, i.e., that the same protein can be assigned to several modules. In addition, we will compare our method with some recently proposed topological methods for module identification (see Chapter 9).

Chapter 7

Weighted core-clustering coefficient for identifying modules

In this chapter, we describe further developments of the method proposed in (Lubovac, et al., 2006) and described in Chapter 6. In SWEMODE (see section 6.2.2), we use a measure called weighted clustering coefficient, which takes into consideration the functional similarity between interacting proteins. Here, we propose further extensions of the method (see section 7.1.1) to investigate if the weighted clustering coefficient should be calculated based on the highest *k*-cores of a graph, instead of the original graph, as proposed earlier. The *k*-core of a graph corresponds to the most densely connected sub-graph in the neighbourhood of a node (Bader and Hogue, 2003). We also introduce the cellular component aspect into the calculation of the weighted core-clustering coefficient, in order to analyse if this inclusion may improve the previous results. The results from applying those new aspects on the CORE data set are presented in section 7.2.1, while the corresponding results based on the new data set from (von Mering, et al., 2002) may be found in section 7.2.2. Another extension of the proposed method is that we allow overlapping modules of the interaction networks, which implies that a protein can be present in multiple functional modules, which is biologically realistic, since proteins may participate in multiple cellular processes and pathways. In section 7.3, we introduce the new type of semantically weighted clustering coefficient that takes into account all three triangle-forming edges, and not only the two of the edges adjacent to a node. This weighted measure is also applied on the CORE data set to identify modules, and the result is compared to the corresponding

result obtained from the original weighted clustering measure used in this work. Finally, a summary and conclusions may be found in section 7.4.

7.1 Materials and methods

As stated earlier, we here employ and analyse further extensions of SWEMODE. according to two important aspects of biological networks: the overlap between modules and the *k*-core aspect. In previous work (Lubovac, et al., 2006) (see Chapter 6), no overlap was allowed between the modules, i.e. proteins were clustered into disjoint modules where each protein could only belong to one module. In this way, modules were treated as isolated functional units, with no possibility to reveal their interconnectivity. However, previous work on the analysis of the yeast filamentation and signalling network indicates that overlapping proteins (Lubovac, et al., 2005) (see Chapter 5) and highly interconnected proteins (Rives and Galitski, 2003) in several cases constitute a part of an intermodule path and may play an important role for intermodule communication.

Next, we introduce the cellular component aspect into the calculation of the weighted core-clustering coefficient, and we also add this aspect to a combined weighted core-clustering coefficient that takes into consideration two of the GO aspects – molecular function and biological process.

Besides introducing an overlap aspect and cellular component information, another extension of SWEMODE considers *k*-cores of the graph. The notion of a core-clustering coefficient has been introduced in previous work (Bader and Hogue, 2003). Here, we propose a weighted counterpart, i.e. a weighted core-clustering coefficient, which takes into consideration semantic weights and topological properties, i.e. information about the highest *k*-cores for a graph. *K*-core decomposition has been proposed earlier for detection of complexes from protein interaction networks (Bader and Hogue, 2003; Tong, et al., 2002). It has also been found recently that proteins that participate in central cores have more vital functions and a higher probability of being evolutionarily conserved than the proteins that participate in more peripheral cores (Wuchty and Almaas, 2005). This also motivates our attempt to improve SWEMODE by including this aspect.

7.1.1 Further development of SWEMODE

The core-clustering coefficient of a node i is defined as the density of the highest k-core of the closed neighbourhood $N[i]$. The highest k-core of a graph is the most densely connected sub-graph (Bader and Hogue, 2003). In this work, we propose a weighted core-clustering coefficient for identifying topologically and functionally cohesive clusters. The proposed weighting scheme, called $core(c^W)$ uses the weighted core-clustering coefficient of node i, which is defined as the weighted clustering coefficient of the highest k-core of the closed neighbourhood $N[i]$ multiplied with the highest k-core number. The use of the weighted core-clustering coefficient, (instead of the weighted clustering coefficient) is advantageous since it amplifies the importance of tightly interconnected regions, while removing many less connected nodes that are usually present in scale-free networks (Bader and Hogue, 2003). The relative weight assigned to node i, based on this measure, is the product of the weighted core-clustering coefficient and the highest k-core number of the immediate neighbourhood of i. By assigning this relative weight to i, the importance of highly interconnected regions is further amplified. There are other functions, such as the density function (Bader and Hogue, 2003), but these are not evaluated here.

SWEMODE has three options concerning traversal of nodes that are considered for inclusion in a module, as described in Chapter 6. In Chapter 6, we applied immediate neighbour search, while we here use depth-first search (DFS), i.e., the protein graph is searched starting from the seed node, which is the highest weighted node, followed by recursively traversing the graph outwards from the seed node, identifying new module members according to the given *NWP* (Node Weight Percentage) criterion. As in previous experiments, SWEMODE was run over a range of 20 *NWP* parameter values (0 to 0.95 in increments of 0.05). At this stage, once a node has been visited and added to the module, it cannot be added to another module (Lubovac, et al., 2006).

However, in the post-processing step, overlap is allowed to some extent. Because we here choose to go further by inspecting the interconnectedness, it is valuable to not only traverse the immediate neighbours but also other indirect neighbours, which may be potential bridges between modules. In a post-processing step, modules that contain less than two members are removed, both before and after applying a so called "fluffing" step. The degree of "fluffing", that is referred to as the "fluff" parameter, corresponds to the weighted cohesiveness of a node, and can vary between 0.0 and 1.0 in increments

of 0.1 (Bader and Hogue, 2003). For every member in the module, its immediate neighbours are added to the module, if they have not been visited and if their neighbourhood weighted cohesiveness is higher that the given fluff threshold *f*. The evaluation is based on 400 module sets. Besides the fluff and *NWP* parameters that were varied, we choose to remove the single-member modules, both before and after increasing the size of modules by fluffing. In (Bader and Hogue, 2003), they used a similar parameter referred to as "haircut". If the algorithm is run with this option, the remaining complexes are 2-cored.

7.2 Results

7.2.1 CORE data set

As in Chapter 6, we here also apply overlap score evaluation of predicted modules against MIPS complexes. The overlap score is defined in Equation 17 on page 90 (for further details, see Chapter 6). We first compared modules obtained from the sub-graph based on highest *k*-cores with the modules obtained from the original PIN (see Figure 25 on page 108). The *k*-core sub-graph is obtained according to the following procedure. For each node *i*, the highest *k*-core number is defined, and all direct neighbours of *i* that have a lower *k*-core are removed from the immediate neighbourhood. Hence, only the neighbours with the same or higher *k*-core number are kept in the neighbourhood. For more details on *k*-core analysis, see section 2.5.4, decomposition, where the outline of the decomposition algorithm may be found in Text box 1 on page 22. When using the original PIN to generate modules, in the post-processing step we allow all direct neighbours from the original PIN to be added to modules (depending on their fluff values), and not only those belonging to the highest *k*-core sub-graph.

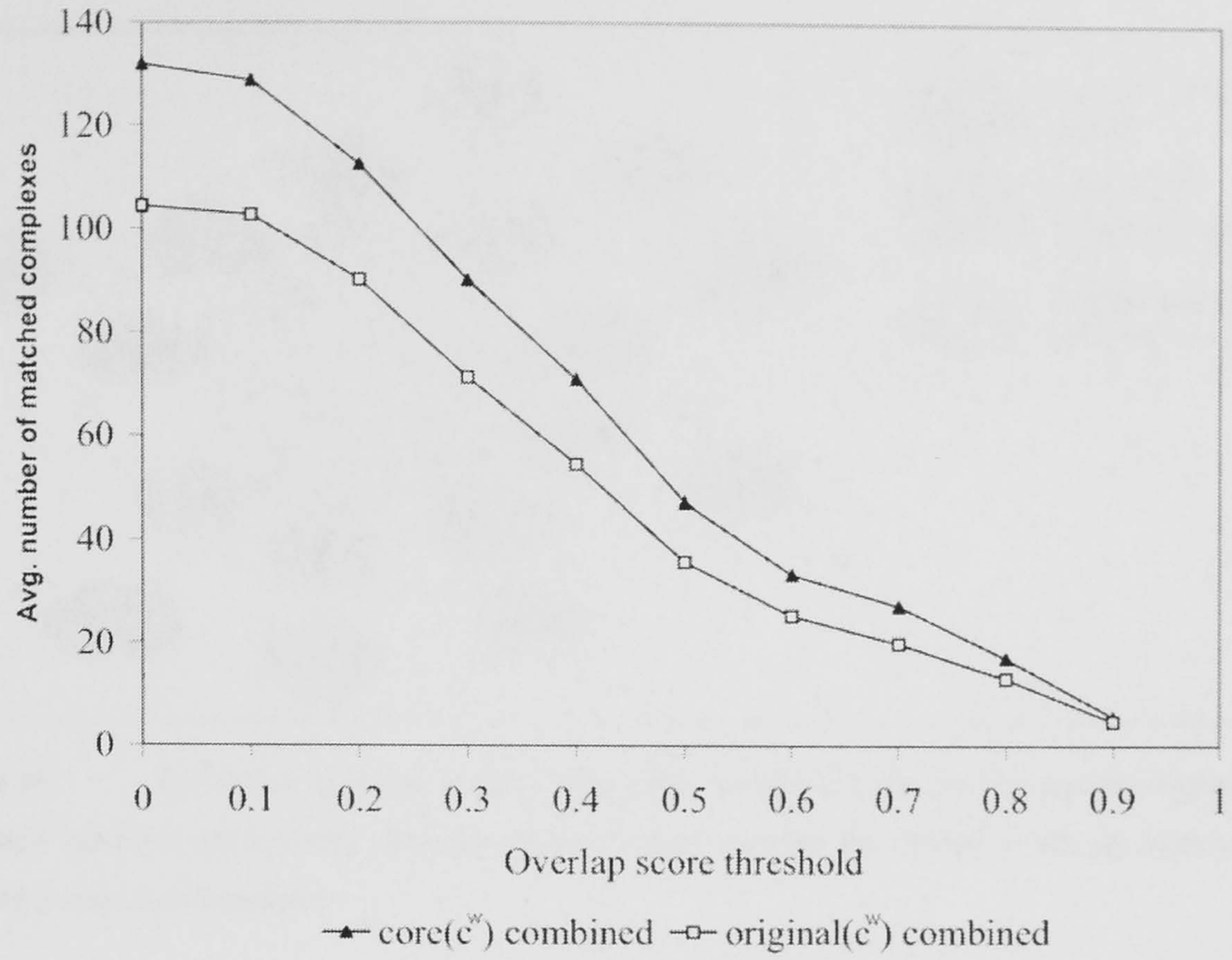


Figure 25: Results from the original PIN versus the *k*-core sub-graph (by using DFS). All three sub-ontologies were used to generate a combined weighted function

It is obvious that the exclusion of the proteins that belong to the original but not to the highest *k*-core graphs increased the overlap with MIPS complexes. Interestingly, at the highest threshold value ($Ol > 0.9$), both networks produce the same results, which indicates that the predicted modules that exactly match the MIPS complexes seem to be very robust and preserved after applying the *k*-core decomposition. This confirms the benefits of using highest *k*-cores of the graph, because, as pointed out in previous work (Wuchty and Almaas, 2005), yeast proteins that participate in the most central cores seem to be evolutionarily conserved and essential to the survival of the organism.

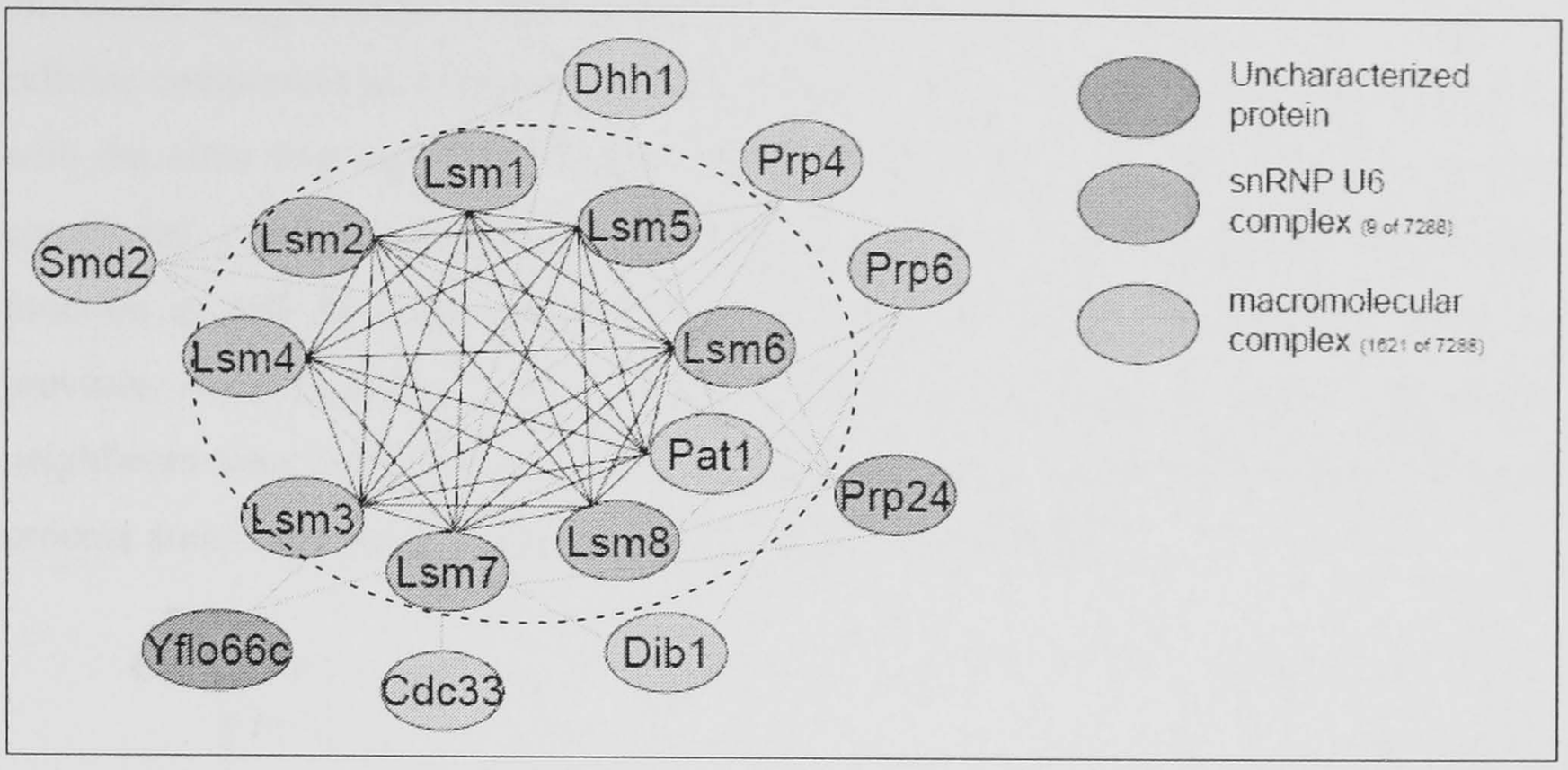


Figure 26: Difference between original sub-graph containing Lsm proteins and the highest *k*-core sub-graph obtained after *k*-core decomposition. Proteins outside the dashed circle are removed after applying *k*-core decomposition

Figure 26 above exemplifies the effect of *k*-core decomposition. It shows the difference between a sub-graph containing Lsm proteins derived from original PIN and the remaining sub-graph after applying *k*-core decomposition. There are nine proteins left in the *k*-core graph (inside the dashed circle), while the remaining eight proteins that are placed outside the marked circle are dismissed. Among the dismissed proteins, one member is uncharacterised (red coloured node in Figure 26) and the remaining proteins (with the exception of Prp24) are directly annotated with term "macromolecular complex" (blue coloured nodes in Figure 26), when using the GO cellular component sub-ontology and SGD as annotation sources. Out of 7288 background proteins, 1621 are directly annotated with this term, meaning that the term is not so specific compared to the corresponding term that is assigned to the *k*-core proteins, "snRNP U6 complex", where the specificity is much higher because eight of totally nine annotated proteins in the background set are left in the *k*-core graph and correspond to a known ribonucleoprotein complex.

Next, we evaluated the effect of using cellular component information to calculate the weighted core-clustering coefficient, and also adding this aspect to a combined weighted core-clustering coefficient that takes into consideration two of the GO aspects – molecular function and biological process. The result of this comparison is shown in Figure 27 on page 110 (when using DFS) and Figure 28 on page 111 (when only the

immediate neighbourhood was considered). As Figure 27 shows, both using GO cellular component as a separate aspect when calculating weights and in combination with the other two aspects generated slightly better results in terms of matched MIPS complexes, compared to the corresponding results generated by using GO molecular function or GO biological process separately. This result is interesting, as we in previous study (Lubovac, et al., 2006) described in Chapter 6, when only direct neighbours were considered for inclusion in the modules, found that the GO biological process annotation was the most suitable for deriving modules.

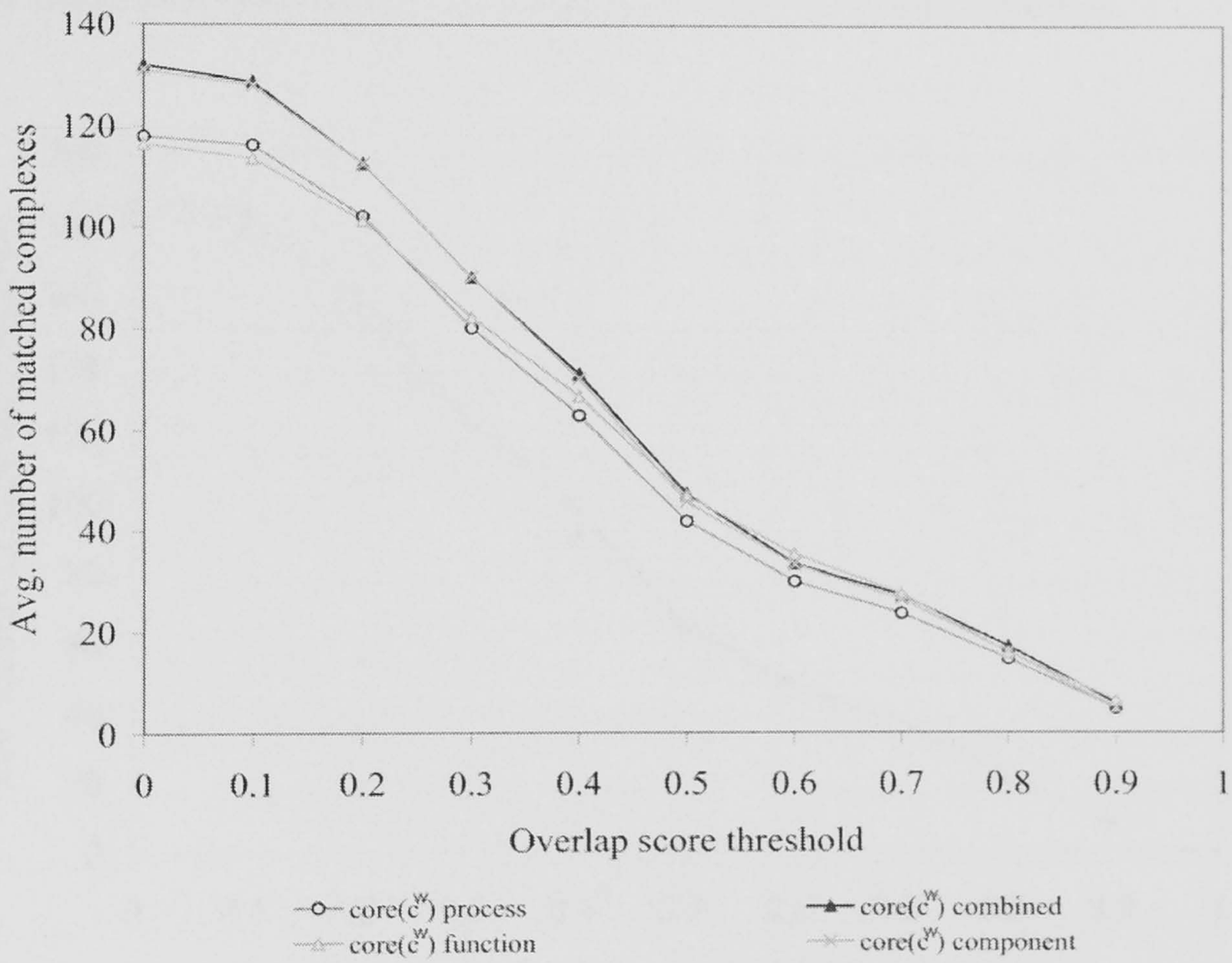


Figure 27: Results from applying separate GO aspects versus the combined measure with SWEMODE when using the DFS option

In Figure 28, we compare the same aspect as in Figure 27 above, but modules have been generated by only considering the immediate neighbourhood for the inclusion of new members in a module. As seen in this figure, where all three separate aspects were compared each other and to the combined aspect, the result from using GO cellular component only is almost the same as the corresponding result when using GO

biological process only, while using combined aspect generates slightly better result. GO molecular function performs worse.

In summary, using only the GO cellular component information in Figure 27 gives results comparable with using all three types of annotations in combination, when DFS is applied for the inclusion of new protein members in modules. Furthermore, considering GO cellular component information when calculating the weight based on the combination of all GO aspects improves the overlap scores slightly, when only direct neighbours are taken into consideration for module inclusion. However, using only GO cellular component is comparable with using GO biological process.

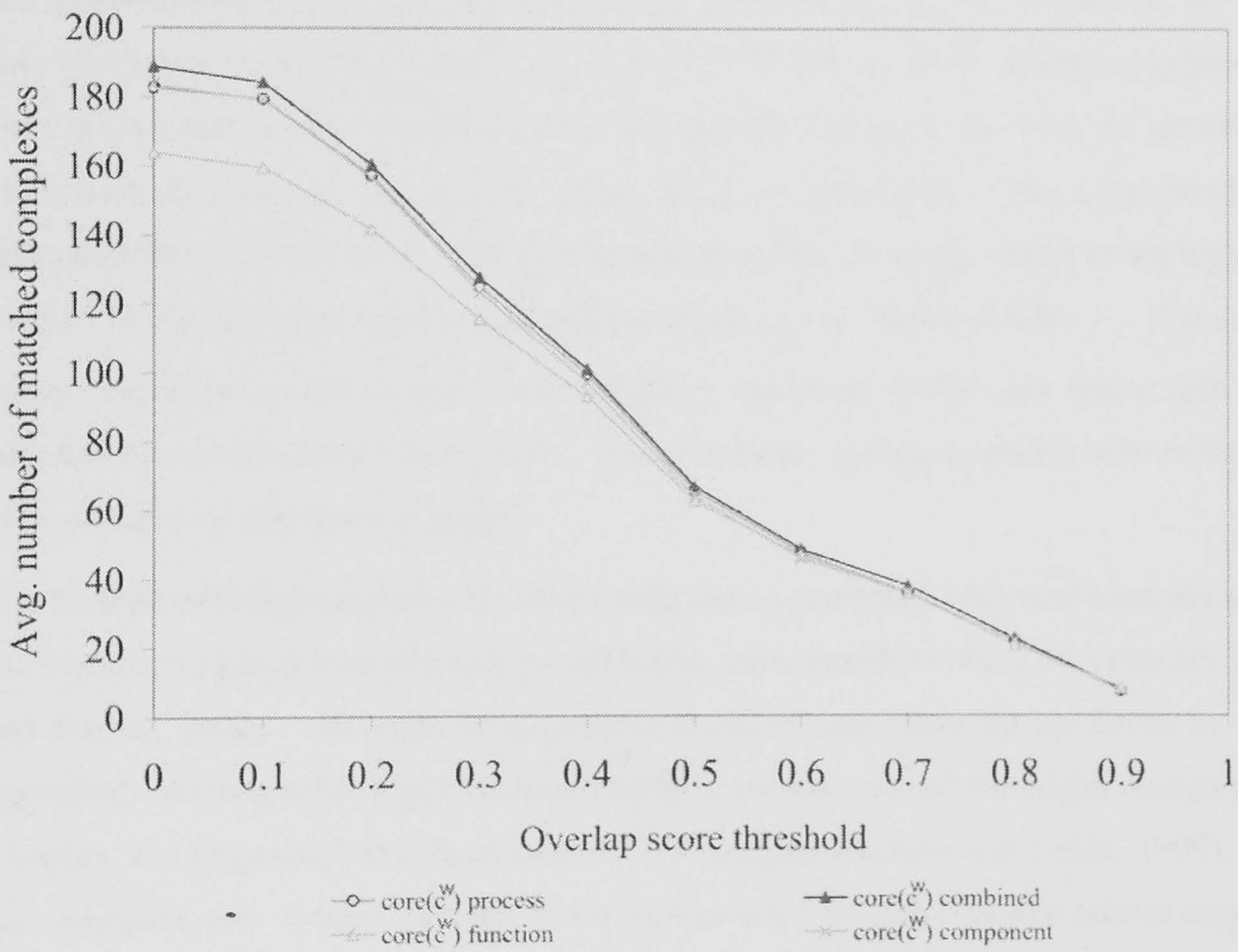


Figure 28: Results from applying separate GO aspects versus combined measure with direct neighbours

This may be explained by the fact that we use different procedures for inclusion of proteins in the modules. When using the DFS option, the algorithm recursively moves outwards from the seed node, identifying indirect neighbours of the seed node whose weights are higher than a certain threshold, which is given as the *NWP* of the seed node.

This indicates that the further we move away from the seed protein, the similarity between the GO terms assigned to the seed protein and the corresponding terms assigned to its indirect neighbours drops faster, when using GO molecular function or GO biological process, compared to the GO cellular component. Simply stated, the seed protein may, for example, be directly or indirectly connected to the neighbours that perform different functional activities, but they may still be a part of the same macromolecular complex (which is described in the GO cellular component sub-ontology). However, when only considering direct neighbours as potential module members, GO biological process is more efficient than GO molecular function in capturing low-level modules.

We further evaluated the importance of introducing module overlap. Generally, we can state that introducing the overlap, i.e. a degree of "fluffing" of the modules, improved our previous results. For example, at the overlap threshold level $Ol > 0.4$, we identified nine more modules on average by using the fluff parameter. The corresponding difference for $Ol > 0.3$ is 12. The best parameter setting, which resulted in the highest number of modules that matched predicted complexes, was obtained with $f > 0$ (i.e. all direct neighbours of the modules with weighted clustering coefficients above zero are added to the module) and $NWP > 0.95$. This parameter setting resulted in 659 modules (471 modules of size three or larger).

The module with highest rank, i.e., the module that is generated with seed proteins with the highest weighted core-clustering coefficient, corresponds to the Lsm complex (He and Parker, 2000). All eight Lsm-proteins (Lsm1-8) are correctly predicted by the algorithm. Among other highly ranked modules, we have found the origin recognition complex, the oligosaccharyl transferase (OST) complex (Knauer and Lehle, 1999), the pore complex, etc. A table with the 10 top ranked and 5 bottom ranked modules may be found in the Table 14 in Appendix C, along with their most significantly shared GO terms describing cellular component.

Finally, we compare the results obtained by using combined measure with DFS option with the corresponding topological measure. As shown in Figure 29 on page 113, the combined measure performs slightly better in terms of the number of matched MIPS complexes. Besides this improvement, we also show in Chapter 9 comparison between the two sets of modules generated with topological approaches and the one proposed in

this work. In Chapter 9, we also discuss some advantages and disadvantages with our approach compared to the described topological approaches.

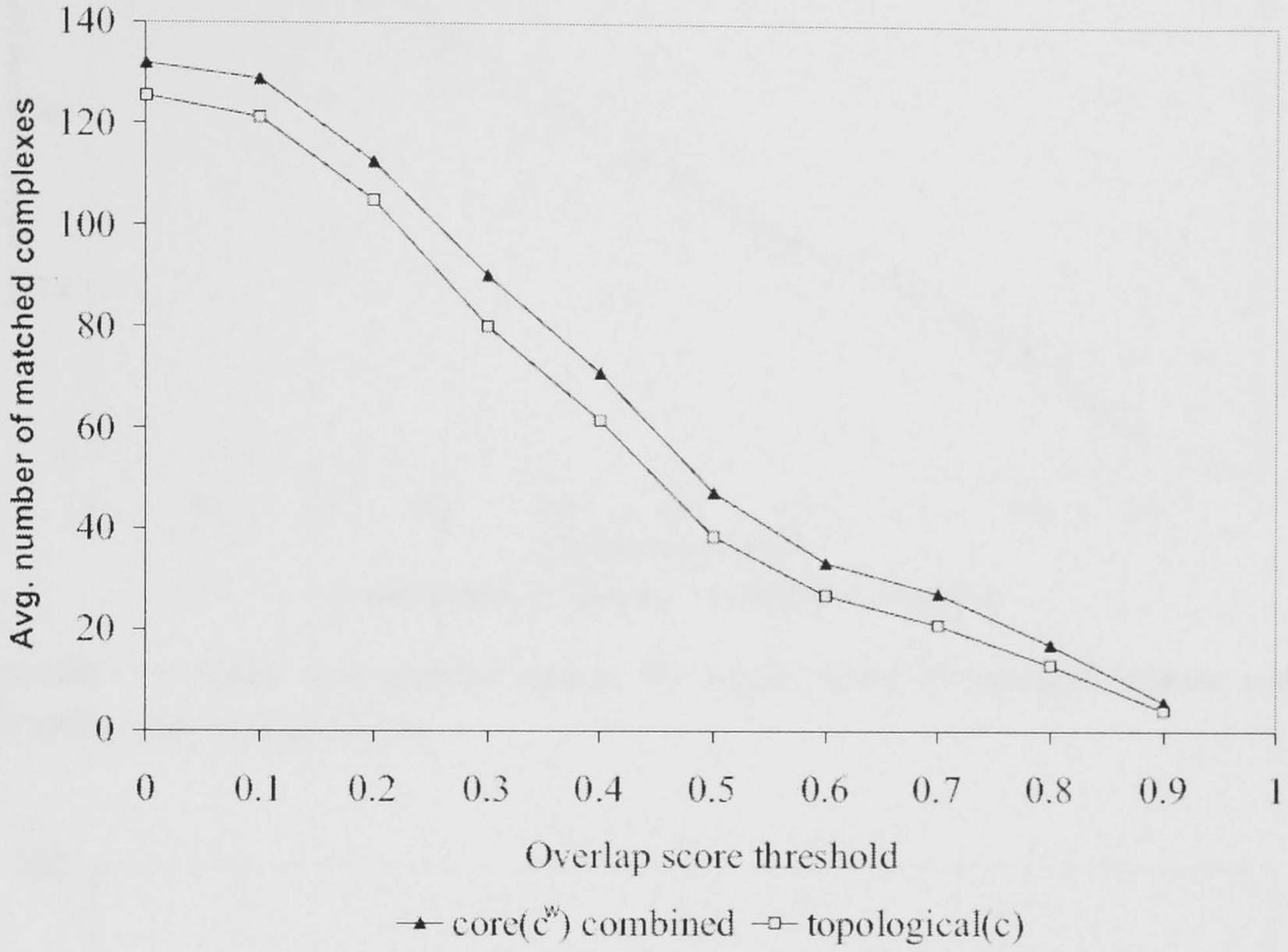


Figure 29: Results from comparing combined measure using DFS option with the topological clustering coefficient based on the same option

7.2.2 Von Mering data set

We performed the same experiments as in section 7.2.2 with the PIN generated from the von Mering data set, to analyse if we get consistent results considering the comparison of different GO aspects. As in the previous section, we applied both immediate neighbour search and depth-first search (DFS). Results in terms of average number of modules that matched MIPS complexes are shown in Figure 30 (DFS) and Figure 31 on page 114 (direct neighbours). As seen in both figures, the GO molecular function aspect performs worse, which is consistent with previous results. Furthermore, using GO cellular component as a separate aspect and the combination of all three aspects generated best results in terms of matched MIPS complexes, while using only GO biological process resulted in slightly lower scores. Hence, the results seem to be consistent with what we found when analysing modules in the PIN generated from the CORE data set (see section 7.2.1).

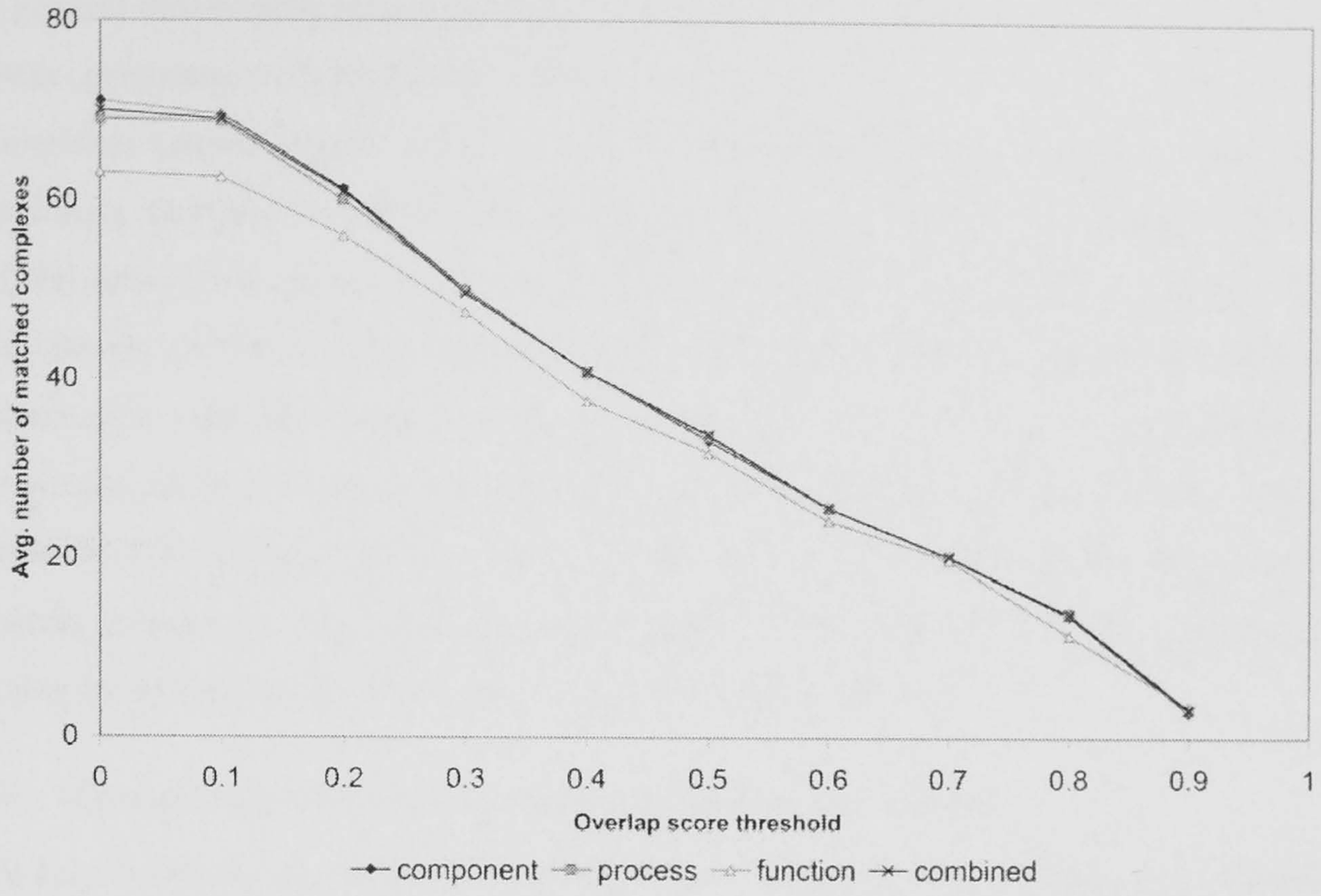


Figure 30: Results from applying separate GO aspects versus the combined measure with SWEMODE when using DFS option

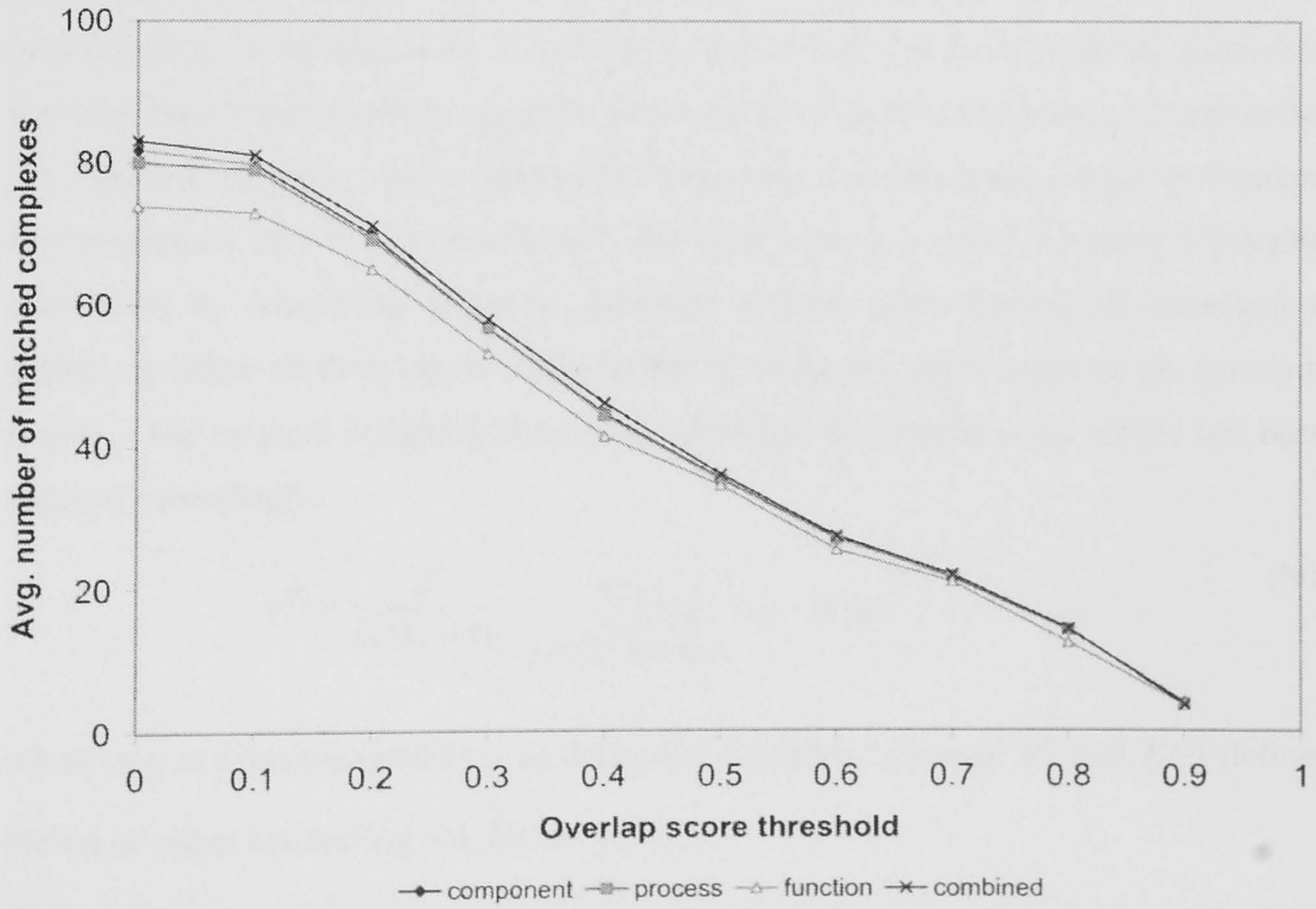


Figure 31: Results from applying separate GO aspects versus combined measure with immediate neighbours

A general observation from the obtained modules based on this data set is that there is a lower percentage of predicted modules with no similarity with any of the MIPS complexes (approximately 24% of modules, based on the average number of matched modules), compared to the corresponding percentage of predicted modules from the CORE data (varies between 29% and 31%, depending on the weighting scheme). This may be due to the fact that the von Mering data set that we have chosen contains only interactions with high level of confidence, meaning that all interactions are confirmed by several methods. Also, the average clustering coefficient for the network obtained from the von Mering data ($C = 0.55$) is higher than the corresponding value for the protein interaction network obtained from the CORE data ($C = 0.34$), showing the higher overall clustering tendency for the von Mering network.

7.3 Comparison with another weighted clustering coefficient

We here introduce an alternative weighted clustering coefficient, adapted from (Onnela, et al., 2005), and apply it in a novel way by using semantic similarity weights.

In Chapter 6, we introduced the notion of combining semantic similarity weights with topological protein-protein interactions by using the weighted clustering coefficient c^{W} (see Equation 15 on page 81). According to this measure, for each triangle formed in the neighbourhood of node i, semantic similarities between i and its two adjacent nodes are summed. However, the weight of the third edge, i.e., semantic similarity between the neighbours of i is not considered. We here employ a novel weighted clustering coefficient by combining semantic similarity weights with topological information, which considers all three edges of the triangles, in an attempt to improve our previous results. The original weighted clustering coefficient by Onnela et al. (2005) has been adapted accordingly:

$$c_i^{W} = \frac{2}{k_i(k_i-1)} \sum_{\forall\, j,h \,|\, \{j,h\} \in K(i)} (ss_{ij} \cdot ss_{ih} \cdot ss_{jh})^{1/3} \qquad (20)$$

where ss_{ij} is semantic similarity, as defined in Equation 5 on page 35, and $K(i)$ defines the set of edges connecting neighbours to node i.

There are several reasons for considering all triangle-forming edges in the analysis of protein interaction networks. As stated earlier, small cliques are more likely to emerge by chance than large ones (Spirin and Mirny, 2003), which is why using semantic

weights of all three edges may help to identify false positives. The weighted clustering coefficient originally proposed by Bader and Hogue (Bader and Hogue, 2003) does not differ from the general clustering coefficient for cliques (see Figure 32 below), which may be a possible disadvantage. In Figure 32, we illustrate the differences between alternative clustering coefficients, applied to a small triangle-formed network. In this figure, the semantic weights between triplets of proteins gradually decrease (from left to right). The values of the general clustering coefficient, denoted as c_i, drop from 1, which is the maximal value, to 0 for the fourth triplet, which misses the link between neighbours adjacent to i. The value of the weighted clustering coefficient that we have adapted from (Barrat, et al., 2004), referred to as c_i^{w2}, is also equal for the first three triangles and drops to 0 for the fourth triangle as $w_{jh} \to 0$. In contrast, the value of the weighted clustering coefficient c_i^{w3} that considers all three edges decreases as $c_i \sim w_{jh}^{1/3}$, tending gradually to 0.

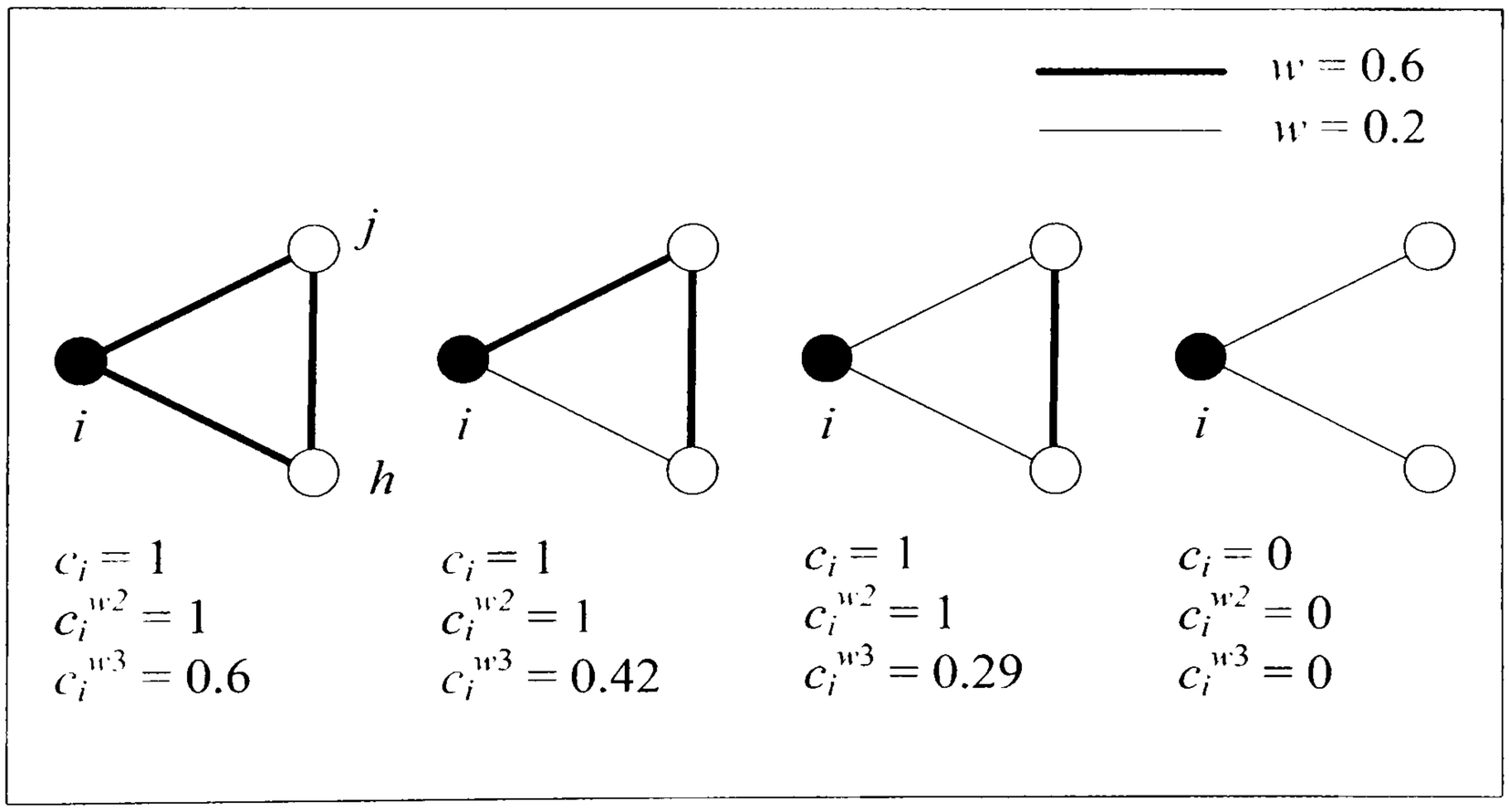


Figure 32: The figure illustrates differences between different clustering coefficients

This is the main reason for why we have employed the approach from (Onnela, et al., 2005) and applied it in a novel way by combining semantic similarity weights with topological information, which considers all three edges of the triangles.

We applied this approach on the CORE data set by using the weighted core function, and combined weights, as in section 7.2.1. The results in terms of the average number

of matched MIPS complexes for different overlap score thresholds are shown in Figure 33 below.

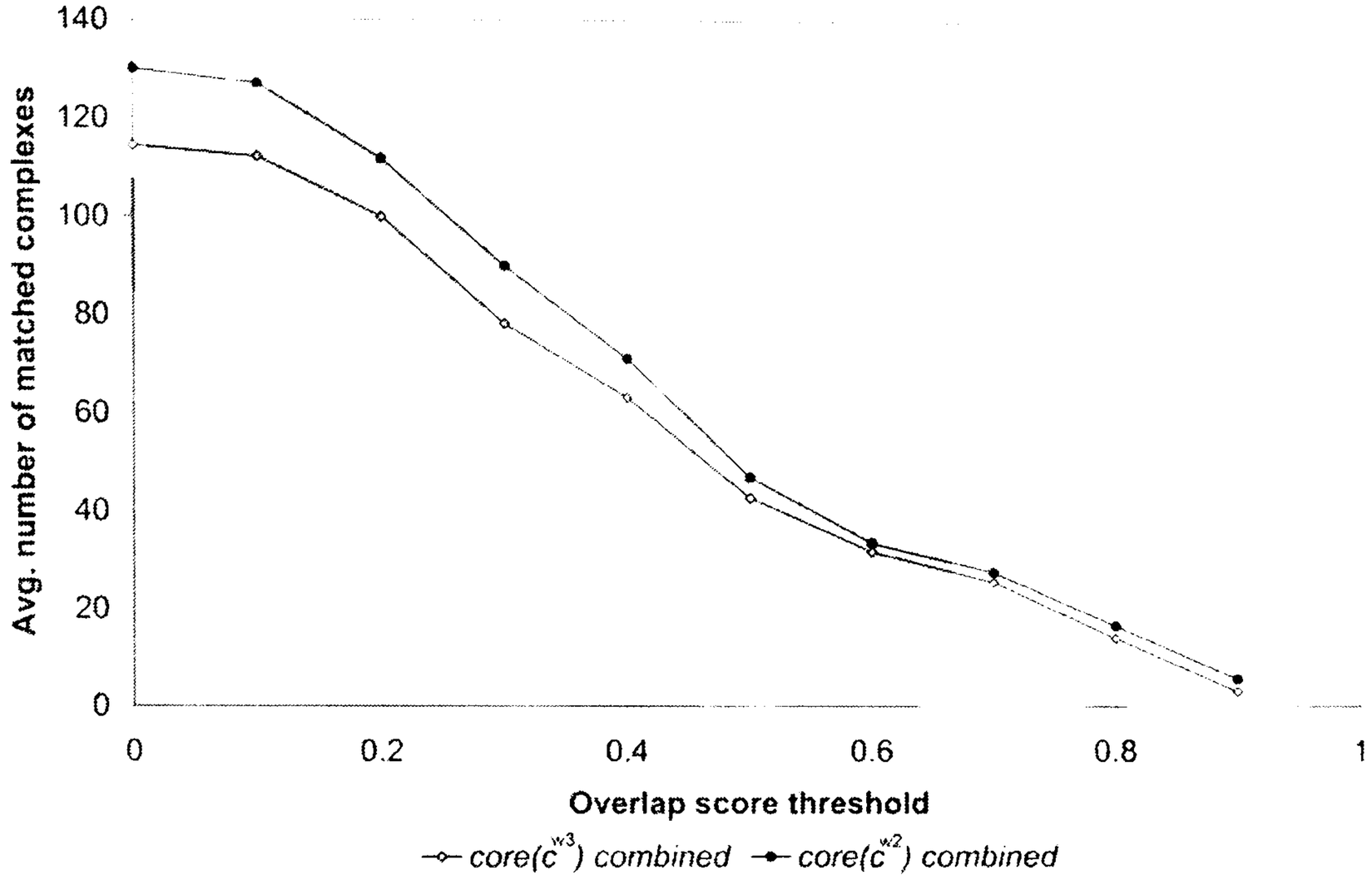


Figure 33: Results of applying SWEMODE by using the core weighting function, based on weighted clustering coefficients c_i^{w2} and c_i^{w3}, and the DFS option

This newly adapted weighted clustering coefficient shows clearly worse performance than the previous one, in terms of the average number of predicted modules that matched MIPS complexes. The weighted clustering coefficient c^{w3} resulted in lower values, and much fewer proteins that can potentially be used to seed the modules, which may have affected results negatively. There are only 901 proteins with weight higher than 0, while calculating weights based on c^{w2} resulted in 1077 potential seed proteins. For example, the protein Kre33, which is used to seed a module containing ribonucleoprotein complex biogenesis and assembly (when using c^{w2} to calculate its weight), while it has weighted clustering coefficient $c^{w3} = 0$.

7.4 Summary and conclusions

We have proposed a method for analysis of protein networks using a measure based on a novel combination of topological and functional information of the proteins (Lubovac,

et al.. 2006). The algorithm takes advantage of this combined measure to identify locally dense regions with high functional similarity. In the evaluation of the method, we found many densely connected regions with high functional homogeneity, in many cases corresponding to sets of proteins that constitute known molecular complexes and some additional interacting proteins which share high functional similarity with the complex, but are not part of it. Together, such sets of interacting proteins form functional modules that control or perform particular cellular functions, without necessarily forming a macromolecular complex. Many of the identified modules correspond to the functional subunits of known complexes. Thus, the method may be used for the prediction of unknown proteins which participate in the identified modules. As indicated by the results, the use of a functionally informed measure to generate modules should imply increased confidence in the predicted function.

We have here demonstrated that restricting the analysis to the highest *k*-core PIN instead of the original PIN resulted in an improved set of modules, with respect to their overlap with known molecular complexes recorded in MIPS.

We were also able to show that using cellular component as a separate aspect when calculating weights, or in combination with the other two aspects, generated slightly better results in terms of matched MIPS complexes, compared to when only two aspects (molecular function and biological process) were included. One of the main reasons accounting for this improvement is the inclusion of the indirect neighbours of the seed proteins in the module prediction step. Proteins that are used as seed modules seem to share more similarity with more distant neighbours when cellular component annotation is used, compared to the two other GO aspects. Seed proteins may, for example be connected, directly or indirectly with neighbours that have different functional activities or are involved in different processes, but they may still be part of the same macromolecular complex (which is described in the GO cellular component sub-ontology).

The results from analysing the von Mering data set also show that the GO molecular function aspect performs worse in module identification, which is consistent with previous results. Furthermore, using the GO cellular component as a separate aspect when calculating weights and the combination of all three aspects generated best results in terms of matched MIPS complexes, while using only GO biological process resulted in slightly lower scores (when using the DFS option). When only considering

immediate neighbours, using the combined aspects performs slightly better than each separate aspect, and GO molecular function still performs worse. Hence, results seem to be consistent with what we found when analysing modules in the PIN generated from the CORE data set.

Finally, we employed a weighted clustering coefficient which considers all three edges of the triangles, because we suspected that this weighting scheme would improve our previous results. However, even if this seemed a reasonable assumption, the results clearly show that considering all three edges may not be a good idea. The weighted clustering coefficient by Onnela et al. (2005) resulted in lower values, and much fewer proteins that are used to potentially seed the modules, which may have affected the results negatively.

Chapter 8

Investigating different features of multi-modular proteins

To generate functional modules as functionally and structurally cohesive structures in PINs is an important step towards reaching the top of the life's complexity pyramid (see Chapter 1) (Oltvai and Barabasi, 2002). However, we need to understand how individual modules communicate and are organised into the higher-order structure(s) of the PIN organisation that underlies cell functionality. To contribute to this understanding, we make an assumption that the proteins that appear in several modules, that we term multi-modular proteins (MMPs), may be useful in building higher-order structure(s) as they may constitute communication points between different modules.

I this chapter, we investigate common properties shared by these proteins, and compare them with the properties of single-modular proteins (SMPs), i.e. proteins that occur in only one module, by analysing three aspects: functional aspect, i.e. annotation of the proteins (see section 8.1), topological aspect that is betweenness centrality of the proteins (see section 8.2), which is used to find topologically important proteins, and lethality (see section 8.3). Furthermore, in section 8.4 we investigate the interconnectivity role of some proteins that are identified as functionally and topologically important.

To identify topologically and functionally important proteins, we calculated the number of module occurrences for each protein across 200 sets of overlapping modules (the fluff parameter was varied between 0 and 1 in increments of 0.1 and the *NHP* parameter was varied between 0 and 0.95 in increments of 0.05). As in Chapter 6, all three GO

aspects were combined into a single weight for each protein. All modules that only contain a single member are removed from further analysis.

For each seed protein, we calculated the number of times each protein appears in different modules in each module set, divided by the number of module sets it appears in. For example, if protein Nup100 is member of 10 modules in one module set, and 20 modules in the another module set, the average number of module occurrences of the protein will be $(10+20)/2=15$.

8.1 GO annotation of multi-modular proteins

8.1.1 CORE data set

We started by analysing annotations with help of SGD GO Term Finder (http://www.yeastgenome.org/help/goTermFinder.html), in order to identify the most significantly shared GO terms among the MMPs with varying number of module occurrence. The sub-ontology "biological process" was chosen. The majority of the most frequent multi-modular proteins (top 10) are annotated with the GO biological process term "cell organization and biogenesis", which has the following GO definition: "the processes involved in the assembly and arrangement of cell structures, including the plasma membrane and any external encapsulation structures such as the cell wall and cell envelope", as described in (Lubovac, et al., 2007). Table 8 on page 122 shows the top ten MMPs, where 80% (highlighted proteins) belong to the above mentioned class. GO Frequency in Table 8 shows the percentage of those proteins that are annotated with the given GO term. The most significantly shared term is obtained by examining the group of proteins to find the GO term to which the highest fraction of the proteins is associated, compared to the number of times that the term is associated with other yeast proteins. The significance (p value) of the shared GO term describing the biological process for the ten most frequent proteins is shown in the last row in Table 8.

In addition, we have repeated the same evaluation procedure by adding proteins with decreasing module frequency to analyse how the annotation statistics is affected by adding those proteins. The summary of those results may be found in Table 15 in Appendix D. The first column shows the statistics for the top 50 protein, where all proteins are present in approximately 2 modules in average. Still, the majority of the proteins share the GO term "cell organization and biogenesis", which is also the most

significant term ($p = 1.3 \cdot 10^{-11}$), and the GO frequency has increased slightly from 80% to 82%. For comparison, 50 random SMPs were evaluated with the same procedure. Here we found that the most significant term that is shared among 96% of those proteins is the GO biological process term "cellular process" ($p = 2.1 \cdot 10^{-5}$), which may not help us to derive any conclusions about the more specific roles of those proteins. Also in this sub-set of proteins, we found that the GO term "cell organization and biogenesis" is shared among proteins, but the GO frequency for this term is 63%, compared to 82% of most frequent MMPs that are annotated with this term.

Proteins	Cdc28	**Nap1**	**Prp43**	Pre1	**Pwp2**	**Sed5**	**Tfp1**	**Nop4**	**Utp7**	**Rpc40**
Module Frequency	4.2	3.9	2.9	2.7	2.7	2.6	2.6	2.6	2.5	2.5
GO biological process	cell organization and biogenesis									
GO frequency	80%									
***p* value**	$3.8 \cdot 10^{-4}$									

Table 8: Annotation statistics for top ten multi-modular proteins

GO term frequency for the most significant terms decreases gradually as we add more proteins with decreasing module frequency. Several non-significant annotation terms appear as we add proteins with decreasing module frequency, meaning that those proteins have more dispersed annotation, while high-frequent MMPs seem to have more consistent annotation dominated by their participation in cellular organisation.

Cdc28, which appears most frequently in modules, is one of five different cyclin-dependent protein kinases (CDKs) in yeast and has a fundamental role in the control of the main events of the yeast cell cycle (Mendenhall and Hodge, 1998). Topologically, it acts as a hub, i.e., it holds together several functionally related clusters in the interaction network (see further section 8.4). In previous work, this protein was suggested to be a part of the intramodule path within the yeast filamentation network, because it had the highest intracluster connectivity, i.e., it was the protein with the highest number of interactions with other members of the same cluster (Rives and Galitski, 2003). It is therefore highly interesting that we have identified this protein as most frequent in our modules, as described in (Lubovac, et al., 2007).

We further evaluated the proteins by analysing their MIPS functional categories (Mewes, et al., 2002), to determine what functional characteristics may be derived by studying proteins based on their module frequency. We observed that proteins involved in cellular organisation (O) appear more frequently among the top 100 MMPs, compared to the random set of SMPs. This result supports our findings based on studying GO biological process annotation, where "cell organization and biogenesis" was the most significant term among multi-modular proteins.

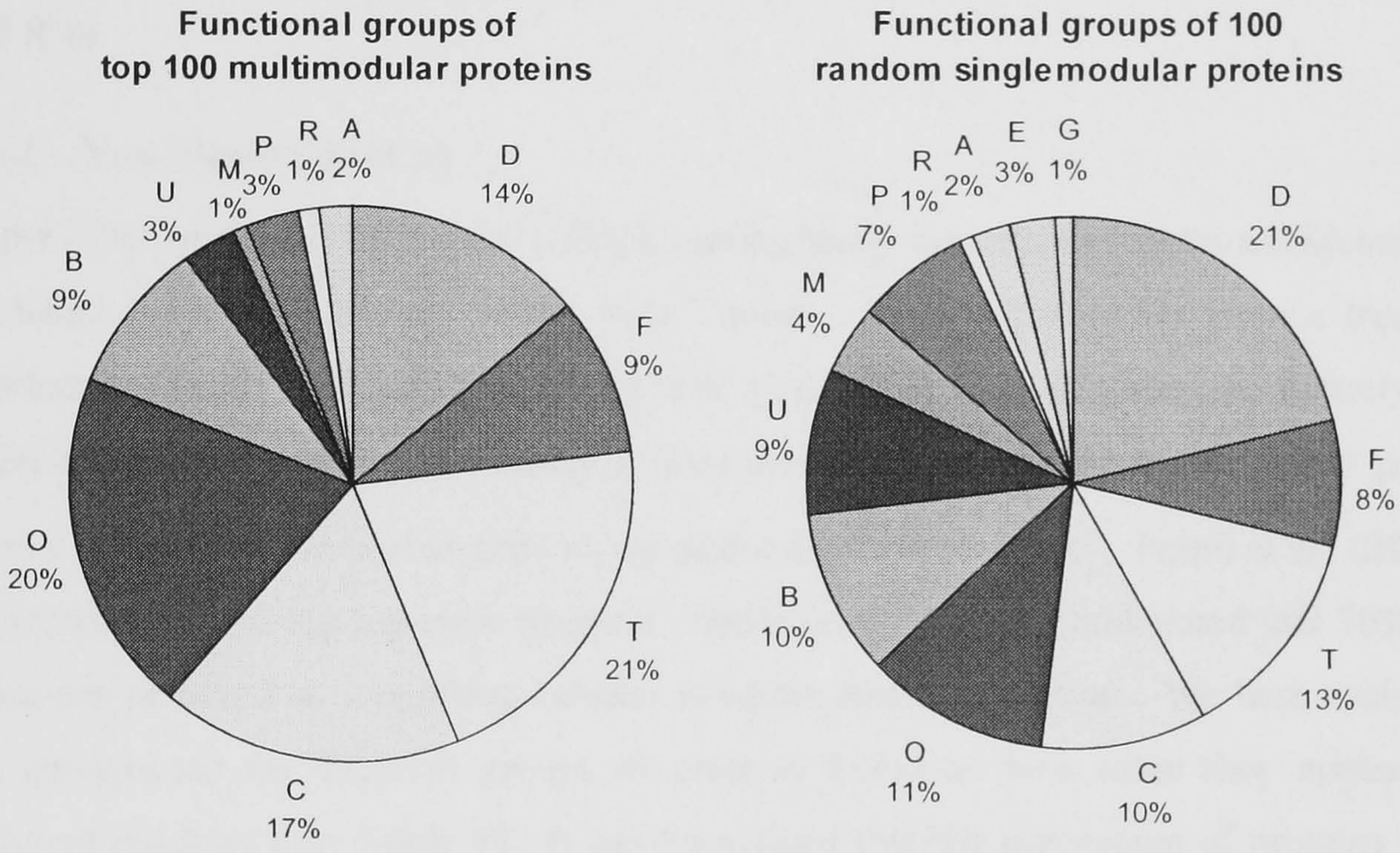


Figure 34: Statistics for MIPS functional categories: D – genome maintenance, T – transcription, F – protein fate, C – cellular fate/organisation, O – cellular organisation, G – amino acid metabolism, M – other metabolism, E – energy production, R – stress and defence, B – transcriptional control, P – translation, A – transport and sensing, U – uncharacterized

We have also found a lower percentage of uncharacterised proteins in the chart that shows the statistics for the 100 most frequent MMPs (see Figure 34a above), while none of the proteins in the top 50 MMPs is uncharacterised (see Figure 43 in Appendix E). This indicates that the more often the protein takes part in the different modules, the higher is the probability that the protein has a defined function. In the same chart (see Figure 34a), the proteins that belong to amino acid metabolism and energy production are absent. By studying Figure 43 in Appendix E, we can conclude that that there is a

high fraction of the proteins belonging to the cellular organisation category in each of the module frequency intervals. To make the charts comparable, we have sorted the proteins in decreasing order of module frequency, and divided them into the four groups of high-frequent proteins, where each group contains 50 proteins (see pie charts in the first row), and four different groups that contain SMPs (see pie charts in the bottom row). The fraction of proteins that belong to the category "cellular organisation" in multi-modular proteins is constantly higher (varies between 18% and 26%) than the fraction of such proteins in the single-modular groups of proteins (varies between 4% and 8%).

8.1.2 Von Mering data set

In previous work by Przulj et al. (2004), topologically important proteins are identified by using the most frequent "bottle neck" nodes. The method starts from a tree of shortest paths for each node v. Such tree consists of n_v nodes that are directly or indirectly connected to v. All nodes w from the tree, such that more than $n_v / 4$ paths from v to other nodes meet at node w, are defined as "bottle necks". Przulj et al., (2004) presented only the top ten most frequent "bottle neck" proteins, and stated that 70% of those are involved in supporting cellular structure and organisation. We here evaluate the annotations for different groups of proteins based on how often they appear in different modules (see Table 9). It can be noticed that the percentage of proteins that are annotated with the chosen terms drops for the proteins with module frequency ≤ 1, with the exception of the term in the last row "primary metabolic process", which is the most common of all presented terms.

Module frequency	≥1.9	≥1.7	≥1.4	≥1.2	>1	≤1
[#] proteins	[50]	[100]	[150]	[200]	[250]	[250]
GO biological process	**GO term frequency** *p* **value**					
ribonucleoprotein complex biogenesis and assembly (5.5%)	42% $9.3 \cdot 10^{-18}$	36% $3.2 \cdot 10^{-18}$	41% $4.9 \cdot 10^{-37}$	40% $1.9 \cdot 10^{-47}$	41% $3.9 \cdot 10^{-64}$	16% $1.1 \cdot 10^{-06}$
cellular component organization and biogenesis (30%)	70% $1.7 \cdot 10^{-06}$	62% $1.2 \cdot 10^{-08}$	65% $1.0 \cdot 10^{-16}$	65% $2.2 \cdot 10^{-22}$	66% $1.9 \cdot 10^{-29}$	56% $3.9 \cdot 10^{-55}$
organelle organization and	50%	45%	51%	50%	53%	35%

biogenesis (17.8%)	$5.8 \cdot 10^{-05}$	$1.0 \cdot 10^{-07}$	$2.1 \cdot 10^{-18}$	$3.1 \cdot 10^{-23}$	$2.1 \cdot 10^{-35}$	$2.2 \cdot 10^{-08}$
RNA metabolic process (14.2%)	44% $8.9 \cdot 10^{-05}$	48% $1.8 \cdot 10^{-13}$	46% $1.3 \cdot 10^{-18}$	45% $5.9 \cdot 10^{-24}$	46% $1.3 \cdot 10^{-31}$	32% $1.1 \cdot 10^{-10}$
primary metabolic process (44%)	74% $4.7 \cdot 10^{-03}$	79% $2.1 \cdot 10^{-10}$	79% $1.1 \cdot 10^{-15}$	81% $8.7 \cdot 10^{-25}$	80% $2.5 \cdot 10^{-30}$	78% $3.8 \cdot 10^{-25}$

Table 9: Statistics for the most significant GO terms based on GO biological process. Module frequency decreases from left to right, and the last column contains a group of proteins that occur in only one module or are not present in any of the modules.

We also present a more systematic comparison between our protein groups, chosen based on their average occurrence in the modules, and the bottle neck proteins (see Table 10). The top 25 proteins obtained by our approach significantly share the term "ribonucleoproteins complex biogenesis and assembly", which is a child term of "cellular component organisation and biogenesis". No significantly shared ontology terms appear in the corresponding set of bottle-neck proteins.

Module freq.	**"bottle necks"**	≥2.1	≥25	≥1.9	≥18	≥1.8	≥14	≥1.7	≥11
[#] proteins		[25]		[50]		[75]		[100]	
GO biological process		**GO term frequency** *p* **value**							
cellular process (64.1%)		-	100% $5.1 \cdot 10^{-3}$	-	98% $2.5 \cdot 10^{-5}$	93% $1.0 \cdot 10^{-6}$	95% $2.0 \cdot 10^{-7}$		94% $9.7 \cdot 10^{-10}$
ribonucleoprotein complex biogenesis and assembly (5.5%)		40% $5.6 \cdot 10^{-5}$	-	42% $9.3 \cdot 10^{-18}$	32% $1.9 \cdot 10^{-6}$	39% $1.7 \cdot 10^{-15}$	27% $1.0 \cdot 10^{-6}$	36% $3.2 \cdot 10^{-18}$	25% $4.8 \cdot 10^{-8}$
cellular component organization and biogenesis (30%)		-	-	70% $1.7 \cdot 10^{-06}$	66% $6.3 \cdot 10^{-5}$	63% $1.5 \cdot 10^{-6}$	61% $9.2 \cdot 10^{-6}$	62% $1.2 \cdot 10^{-8}$	63% $4.2 \cdot 10^{-9}$
organelle organization and biogenesis (17.8%)		-	-	50% $5.8 \cdot 10^{-5}$	46% $1.6 \cdot 10^{-3}$	48% $5.7 \cdot 10^{-7}$	43% $2.2 \cdot 10^{-4}$	45% $1.0 \cdot 10^{-7}$	43% $2.0 \cdot 10^{-6}$
cellular metabolic process (46.6%)		-	-	-	76% $8.6 \cdot 10^{-3}$	79% $3.6 \cdot 10^{-6}$	79% $5.9 \cdot 10^{-6}$	81% $4.7 \cdot 10^{-10}$	77% $2.4 \cdot 10^{-7}$
RNA metabolic		-	-	48%	-	52%	35%	48%	32%

process (14.2%)			$1.8 \cdot 10^{-13}$		$2.8 \cdot 10^{-12}$	$3.4 \cdot 10^{-3}$	$1.8 \cdot 10^{-13}$	$2.4 \cdot 10^{-3}$
primary metabolic process (44%)	-	-	79% $2.1 \cdot 10^{-10}$	-	79% $2.7 \cdot 10^{-7}$	73% $1.2 \cdot 10^{-4}$	79% $2.1 \cdot 10^{-10}$	73% $2.0 \cdot 10^{-6}$

Table 10: Comparison between top 100 most frequent multi-modular proteins and most frequent "bottle neck" proteins, identified by Przulj et al. (2003)

8.2 Betweenness centrality

Betweenness centrality has been applied in the context of social networks, to measure the centrality and influence of a person or a group (Freeman, 1979). The betweenness centrality of a node v, is originally defined by Freeman (1977) as the number of shortest paths (also called *geodesics*) between other nodes that pass through v and it is given by:

$$C_B(v) = \sum_{i,j \in V : i \neq j, i \neq v, j \neq v} \frac{g_{ivj}}{g_{ij}} \tag{21}$$

where g_{ivj} is the number of shortest path linking i and j that contain v, and g_{ij} is the total number of shortest path between i and j. High-betweenness nodes occur on large number of non-redundant shortest paths between other nodes. If a node with high betweenness centrality is removed, it may disconnect different part of the network completely. Thus, such nodes may be thought of as potential bridges between modules in network and have most influence on the information transfer.

8.2.1 CORE data set

We started by investigating general properties of the data set by studying the relation between degree and betweenness centrality. Figure 35 on page 127 shows degree k versus betweenness centrality plotted on algorithmic scale. The few highly connected nodes (hubs) in the PIN must have high-betweenness values because there are many nodes directly and exclusively connected to these hubs and the shortest path between these nodes goes through these hubs. However, the low-connectivity nodes also exhibited a wide range of betweenness values in the yeast PIN.

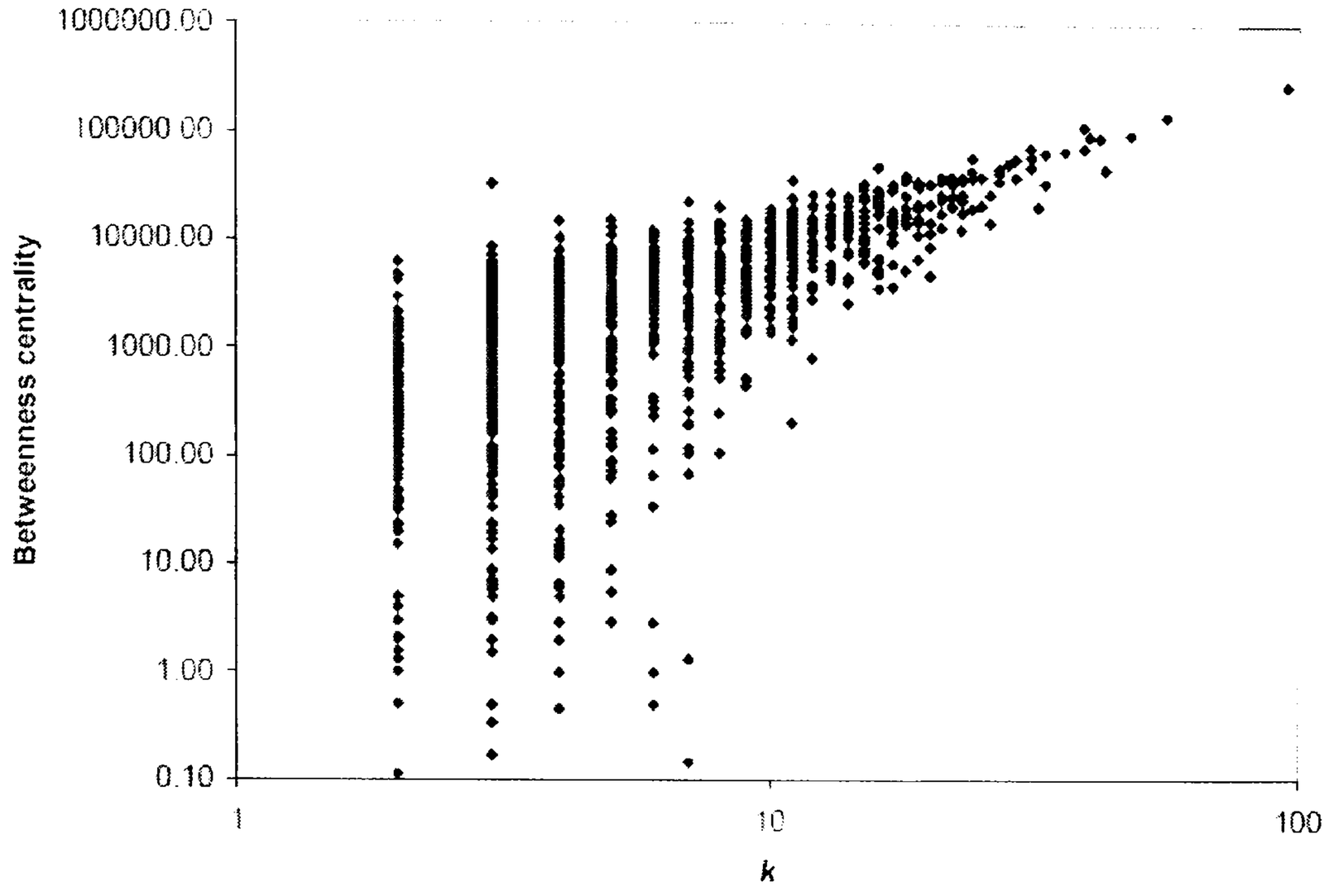


Figure 35: Degree (k) versus betweenness centrality plotted on algorithmic scale

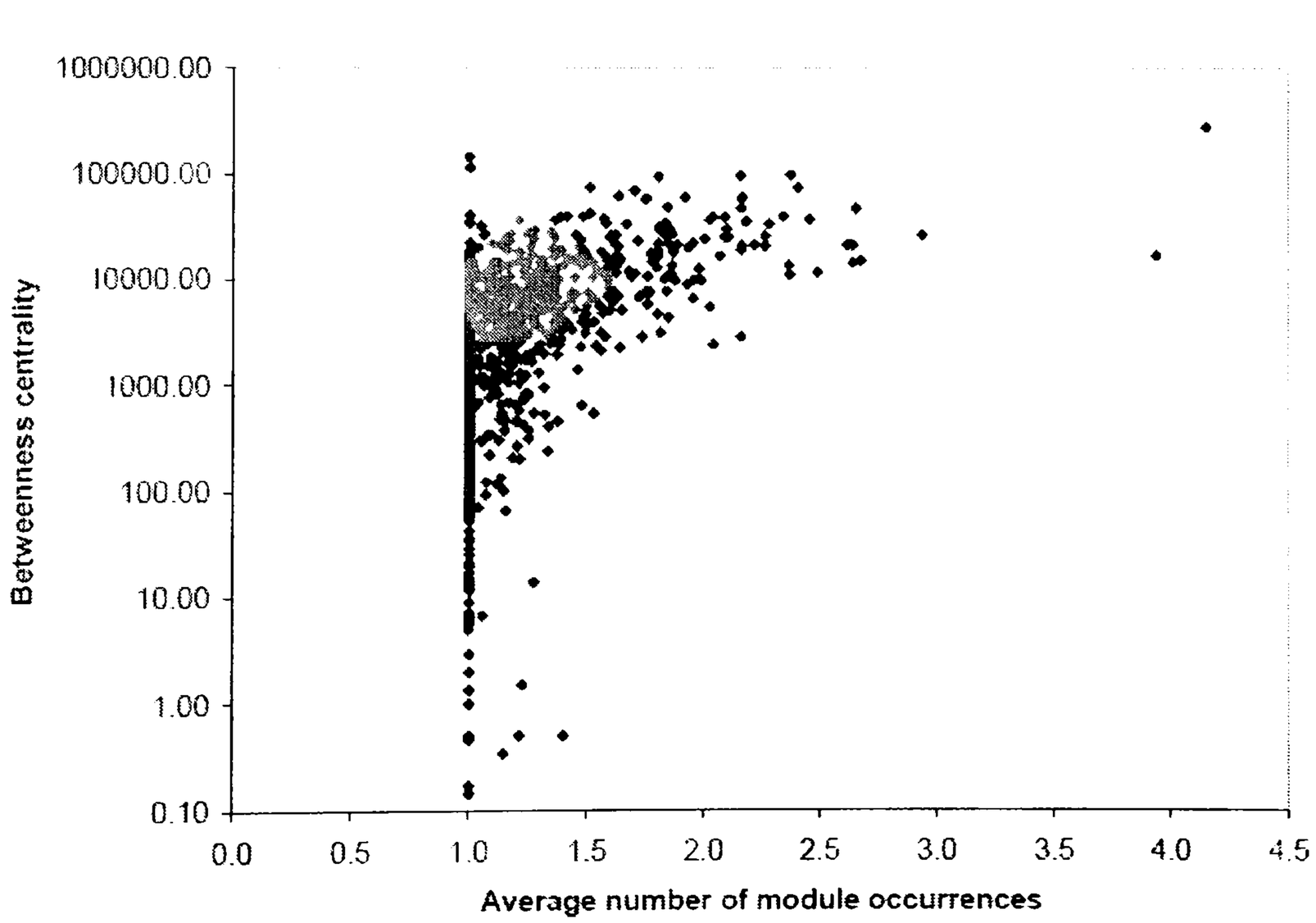


Figure 36: Average number of module occurrences versus betweenness centrality plotted on algorithmic scale

In Figure 36, node betweenness centrality is plotted as a function of average number of module occurrences. We can notice that all proteins with average module frequency ≥ 2 have considerably high betweenness values. However, the single-modular nodes also exhibited a wide range of betweenness values in the yeast PIN.

8.2.2 Von Mering data set

We repeated the same experiment for the von Mering data set. In Figure 37 below, betweenness is plotted as a function of degree k. Here, we could not use any characteristic degree k or any interval of k values to denote the importance of nodes (based on the betweenness).

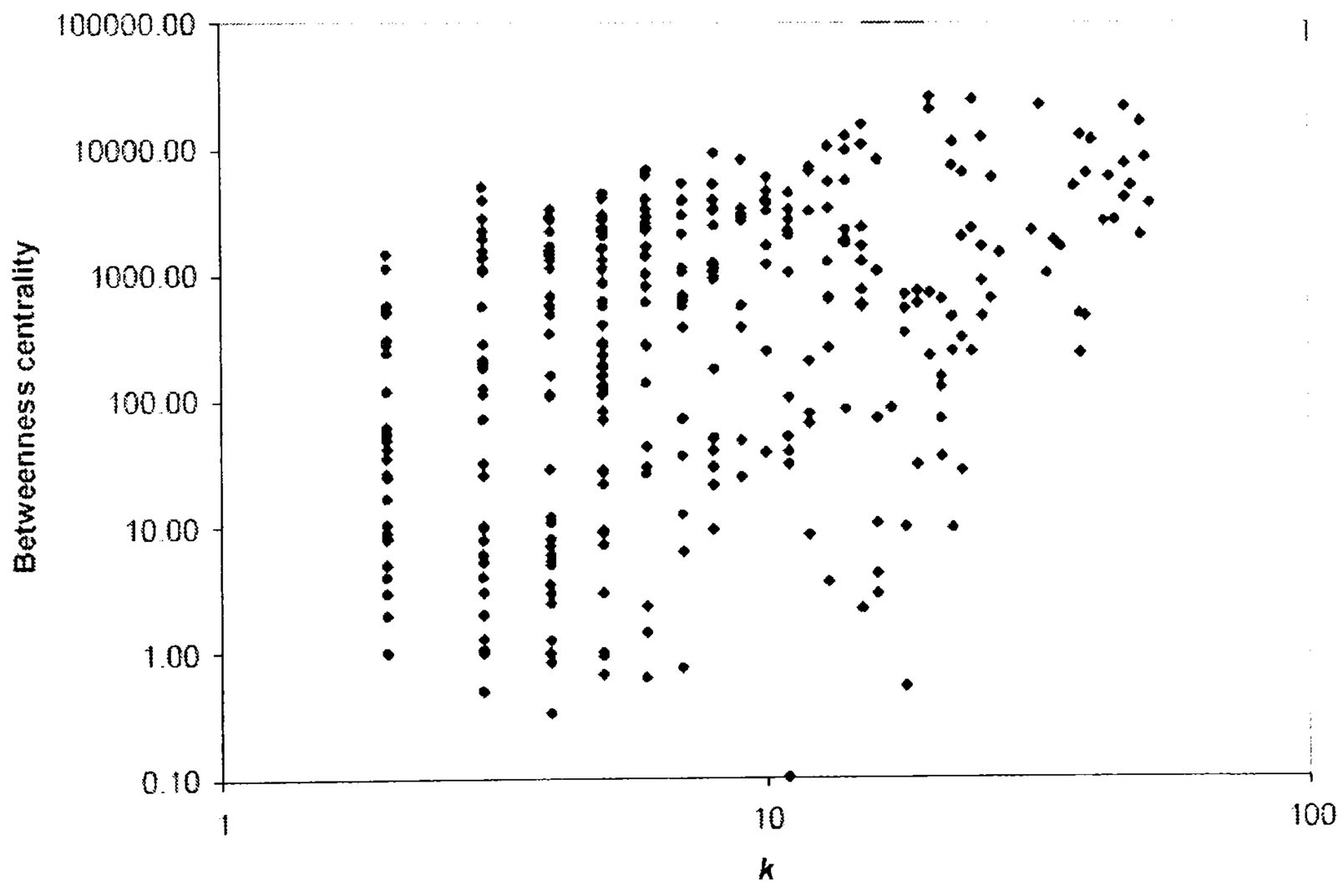


Figure 37: Degree (k) versus betweenness centrality plotted on algorithmic scale

Also in Figure 38 on page 129, besides the most frequent multi-modular proteins (MMPs) that have high betweenness values, there is a wide range of betweenness centrality values for single-modular proteins (SMPs) as well. However, modular frequency seems to be a better indicator of node importance, in terms of betweenness centrality.

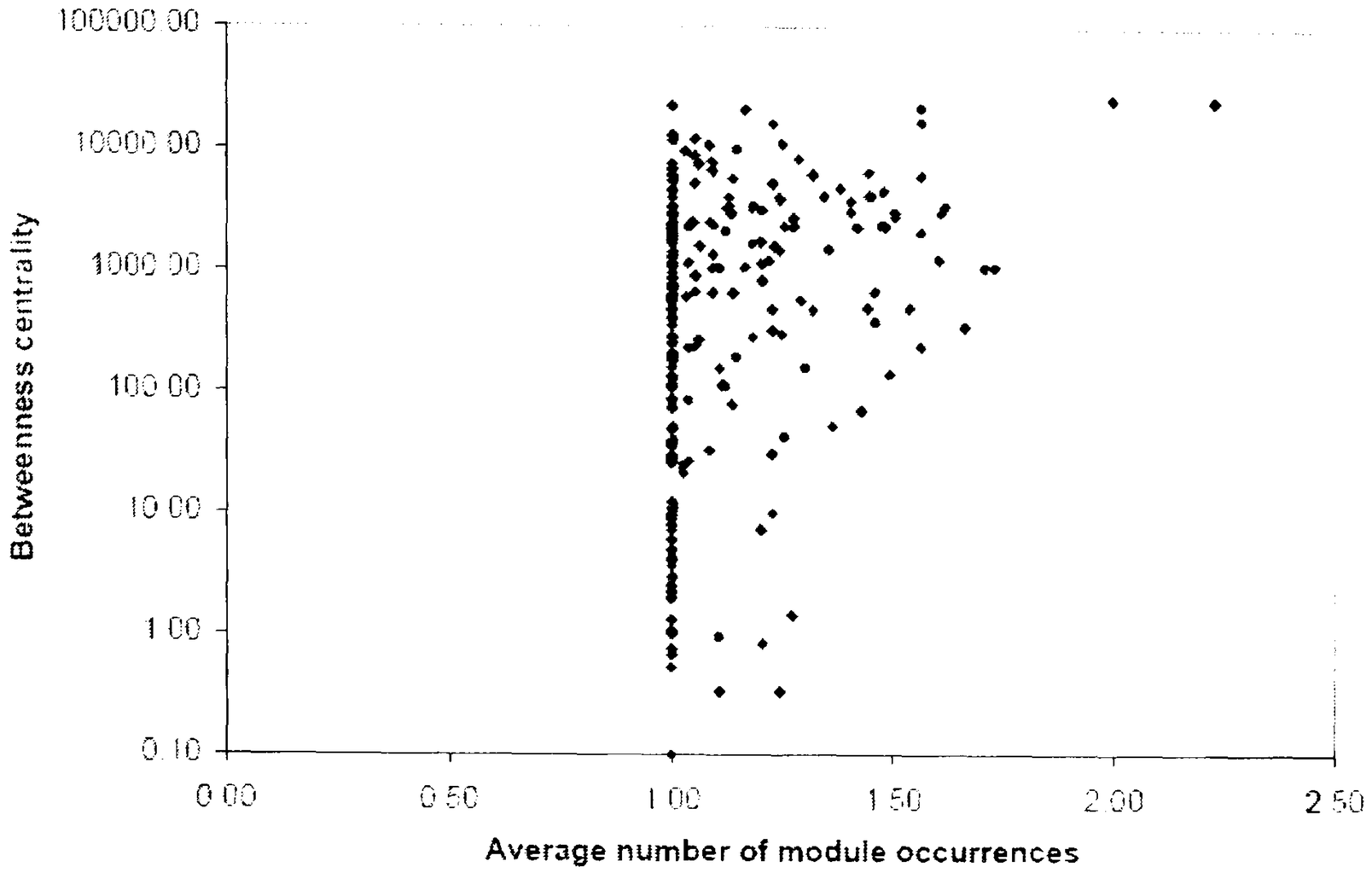


Figure 38: Average number of module occurrences versus betweenness centrality plotted on algorithmic scale

8.3 Lethality

There are 1015 lethal proteins obtained from manually curated MIPS database. The list of MMPs and SMPs observed across modules in both data sets was compared to the list of lethal proteins. The results from this comparison are presented in Tables 11 and 12.

In the CORE data set, we found 222 lethal proteins among the multi-modular proteins (MMPs). This corresponds to 46.3 %, as there are 480 frequently occurring proteins in total. The corresponding percentage for MMPs derived from modules in the von Mering data set is 68.7. as there are 57 lethal proteins among the 83 MMPs (see Table 11 below).

	No. of MMP	No. of lethal proteins	Percentage
CORE	480	222	46.3 %
Von Mering	83	57	68.7 %

Table 11: Lethality among multi-modular proteins (MMPs) across both data sets

We made the same comparison for single-modular proteins (SMPs) across the modules based on both data sets. In the CORE data set, we found 173 lethal proteins among the

SMPs, which corresponds to 34.5 %, as there are 502 SMPs in total (see Table 12). The corresponding percentage for the fraction of lethality in SMPs derived from modules in the von Mering data set is 54.5, as there are 116 lethal proteins among the 213 SMPs, as shown in Table 12 below.

	No. of SMP	No. of lethal proteins	Percentage
CORE	502	173	34.5 %
Von Mering	213	116	54.5 %

Table 12: Lethality among single-modular proteins (SMPs) across both data sets

In both cases, the difference is statistically significant at a 95% confidence level, meaning that there is a significantly larger proportion of lethal proteins, also referred to as important proteins, among multi-modular proteins. These results are obtained by performing a z-test for the differences between the two proportions (z = 3.8 in the CORE data set, and z = 2.2 in the von Mering data set).

8.4 Modular interconnectivity

Figure 39 on page 131 shows the result from an example run from module-identifying method, where Cdc28 was predicted as taking part in six modules matching MIPS complexes. Cdc28 is a cyclin-dependent kinase and it is believed to be a key regulator of the cell-division cycle. In this example, it is connected to several proteins from Origin Recognition Complex (ORC), which is involved in DNA replication. Cdc28 is also connected to actin cytoskeleton-associated complex, which is reorganised in accordance to cell-cycle progression. This process is according to previous study believed to be controlled, directly or indirectly, by Cdc28 (Tang and Cai, 1996). Furthermore, there is an important connection between Cdc28 and proteasome complex. The central role of this complex is to direct a cell to proceed with the decision to replicate itself. In yeast cells a critical trigger for cell replication is degradation of Sic1, which is a protein that inhibits the chemical activity of Cdc28. After eliminating the biochemical Sic1 "brake" due to the action of SCF and the proteasome, the kinase is then free to trigger the progress toward DNA replication and associated events of cell replication.

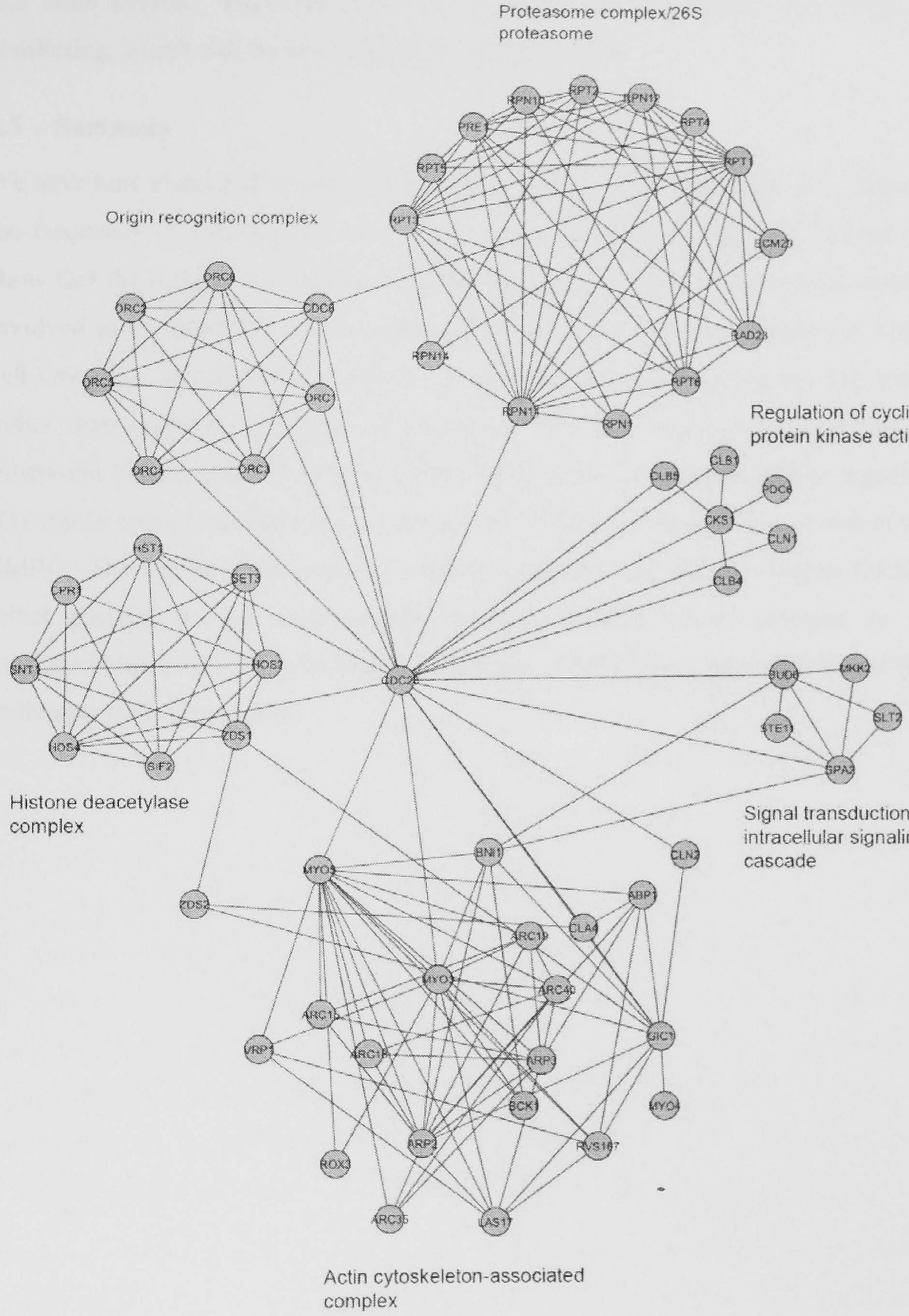


Figure 39: Modular network involving modules in which Cdc28

This is a clear example of the network involving hub that interconnects several functional modules. This example is supported by several topological and functional features, such as average number of occurrences in modules, betweenness centrality,

and node degree. However, there are several examples where those features are conflicting, which will be interesting to evaluate in future.

8.5 Summary

We have here identified topologically and functionally important proteins by calculating the frequency of each protein across 200 sets of overlapping modules. Initial results show that the majority of frequently appearing proteins that connect several modules is involved in the assembly and arrangement of cell structures, such as the cell wall and cell envelope, which indicates that they are involved in supporting the cell structure rather than signal transduction for example. We also observed by studying MIPS functional classes of the MMPs and SMPs that proteins involved in cellular organisation (O) appear more frequently among the top 100 MMPs, compared to the random sets of SMPs. The results from studying lethality show the significantly higher fraction of lethal proteins among multi-modular proteins (MMP), when compared to single modular proteins (SMP) reflecting the tendency of MMP to be more lethal, and hereby indicating their essentiality.

Chapter 9

Comparison with other module definitions

Recently, a topology-based method for detecting modules from a PPI network has been proposed by Luo and Scheuerman (2006) and further analysed in (Luo, et al., 2007). The algorithm uses a new module notion based on the degree definition of the sub-graphs. The approach is based solely on topological properties of the protein sub-graph. It is applied on the same CORE data set that we have used here. A total of 99 modules were detected in (Luo and Scheuermann, 2006). A new agglomerative algorithm was developed to identify modules from the network by combining the new module definition with the relative edge order generated by the Girvan-Newman algorithm. A JAVA program, MoNet, was developed to implement the algorithm Luo et al. (2007). Applying MoNet to the yeast core protein interaction network from the database of interacting proteins (DIP) identified 86 simple modules with sizes larger than 3 proteins. For convenience, those modules will be referred to as MoNet modules.

We have evaluated the MoNet modules with the overlap score threshold, and compared them with our modules generated across approximately 400 different parameter settings, and found that our modules show higher agreement with MIPS complexes (see Figure 40 on page 134). This comparison also indicates that introducing knowledge in terms of semantic similarity into the network topology seems to be advantageous over using only topology information. Furthermore, this method produces one single partition of the network, which does not seem biologically plausible, as many proteins may be involved in different processes.

Comparison with other module definitions

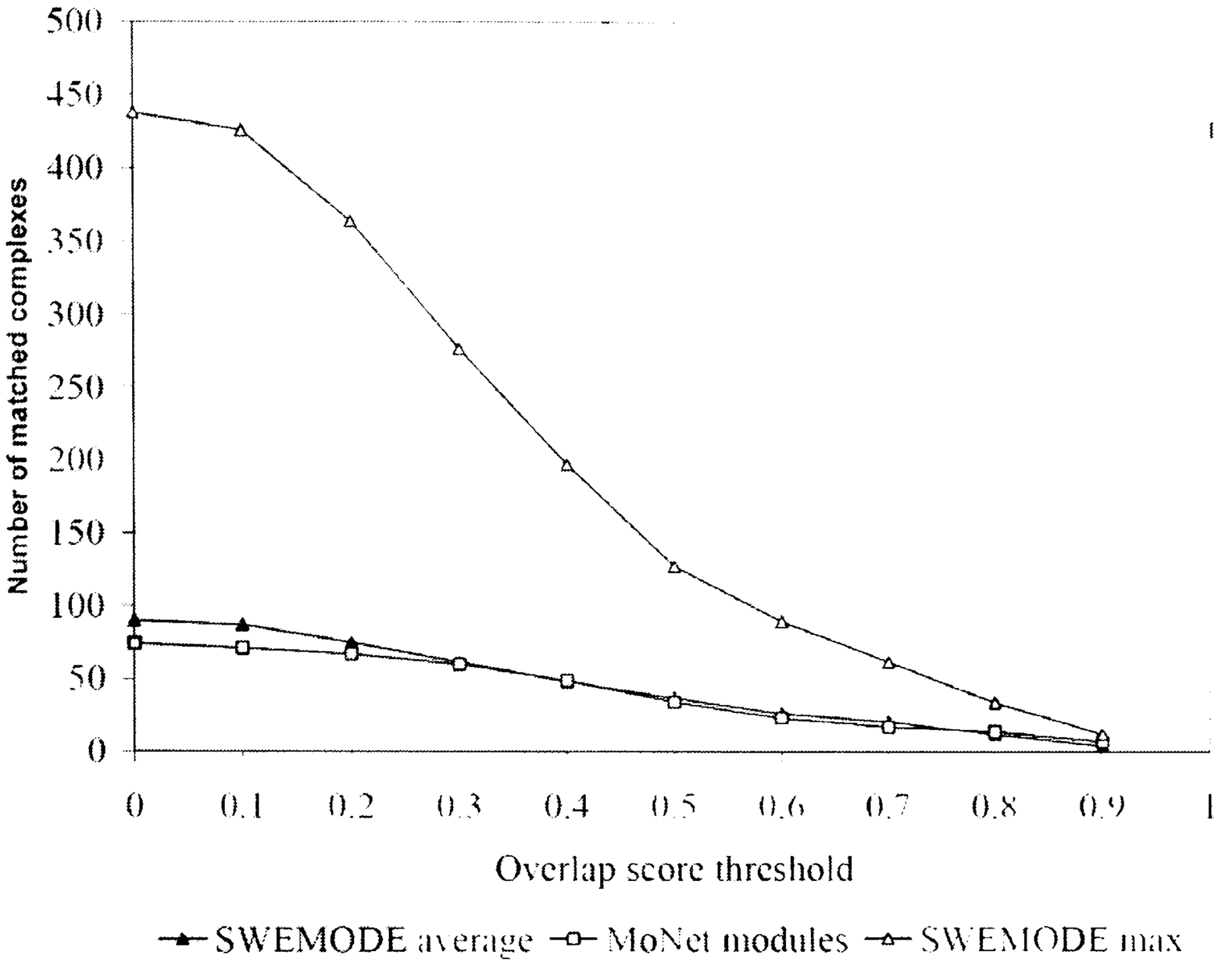


Figure 40: Comparison between MoNet and SWEMODE modules

We also compared our SWEMODE modules obtained from von Mering data with the modules derived in (Przulj, et al., 2004), based on HCS (Highly Connected Subgraphs) clustering algorithm (Hartuv and Shamir, 2000). This method aims to find disjoint subsets (clusters) that should satisfy following criteria: homogeneity – members of the same cluster are highly similar to each other; and separation: members of different clusters have low similarity to each other. The modules generated with SWEMODE showed also here higher overlap with MIPS complexes (see Figure 41 on page 135). A more detailed analysis shows that both algorithms resulted in 39 identical modules. However, as HCS only discern the complexes that are highly interconnected, it discards many clusters that correspond to known complexes. Another disadvantage of both methods that are here compared to SWEMODE is that they do not allow any overlap between modules, i.e. they produce disjoint clusters.

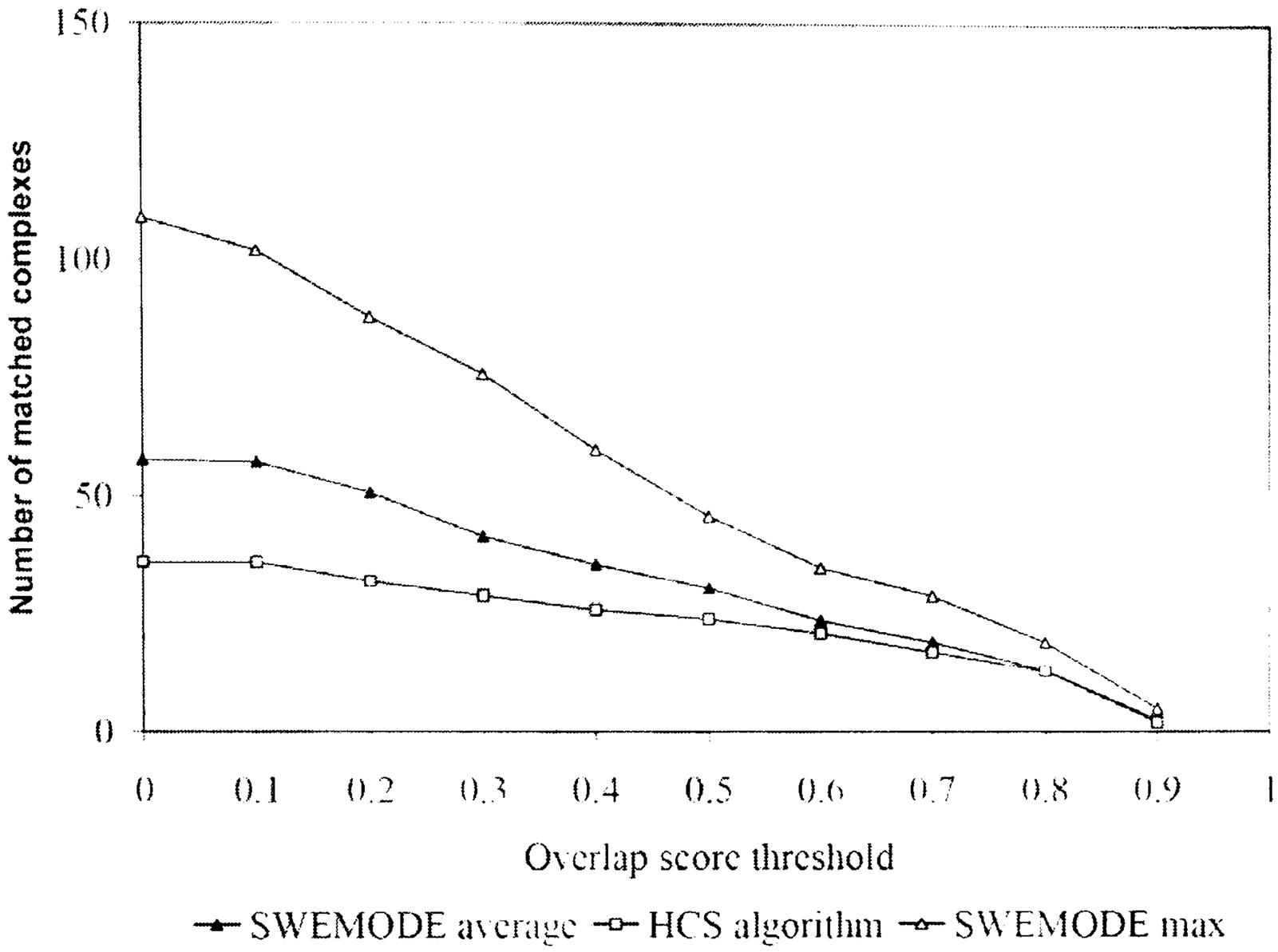


Figure 41: Comparison between SWEMODE modules and modules generated with HCS clustering algorithm

Chapter 10

Conclusions and future work

10.1 Conclusions

In the Introduction of this thesis (Chapter 1, section 1.3), we have delineated 7 contributions that have resulted from this work. It is proposed that those contributions constitute a substantial addition to the body of knowledge in systems biology, which is particularly concerned with protein interaction network analysis. This has been demonstrated throughout the thesis in the following manner:

1 **Comparison of the three most common semantic similarity measures extends previous work by evaluating their performance on the data sets with different degrees of clustering, which is an important property for the purpose of this work**. This adds to our general understanding of the properties and usefulness of these measures in large-scale PIN analysis. The evaluation of the proposal is performed in Chapter 4. We have been able to show that overall accuracy is higher for the data sets with higher degrees of clustering. The results from the chapter show that the semantic similarity measure by Lin performed slightly better than the other two tested measures, when tested on the data set with highest degree of clustering, and worse with the corresponding data set with the lowest degree of clustering. Besides two already stated advantages of the Lin measure, i.e. the fact that it uses the information of the shared parent term, together with the specific term, and that its value varies between 1 and 0, the results from Chapter 4 also support the decision to choose the Lin measure for further experiments, as it shows slightly better accuracy than the other two measures when applied on the data set that has the highest clustering degree.

2 **Development and evaluation of a post processing PIN clustering approach that produces an overlapping modular structure by merging the clusters based on topology information (mutual clustering profiles) with the clusters based on semantic similarity profiles.** The fact that modules may contain overlapping proteins and are based on the combined information reveals additional knowledge that is missed by methods that produce disjoint clusters based solely on topology information. This has been demonstrated in Chapter 5. As discussed in Chapter 5, this approach generates overlapping clusters, where one protein may be a member of several clusters, which is biologically realistic, unlike most clustering approaches that produce disjoint clusters. Another advantage of this approach compared to traditional hierarchical clustering approaches is that the cut-off for the topological clustering is chosen based on the best agreement with domain knowledge. This is a significant extension of the previous work, where cut-off is often based on visual inspection, like in the work by Rives and Galitski (2003) for example. Further results from the analysis of both the filamentation and signalling networks discussed in Chapter 5 show that proteins that are assigned to more than one module in several cases play important roles in intermodule communication.

3 **Development and evaluation of integrated protein-similarity measures that combine semantic weights based on protein annotations and information based on PIN-topology**. In Chapter 6, two metrics that arise from this are described, namely weighted clustering coefficient, and weighted average nearest neighbours degree. The measures are used to probe the properties of the weighted network. For both metrics, we demonstrate their use in analysing the properties of PINs. The comparison between the weighted clustering measure that uses semantic weights, and its topological analogue, described in Chapter 6, shows that the weighted measure generally has higher values than its topological counterpart. Those measures constitute a base for deriving various weighted schemes, which are in turn used to identify modules in Chapters 6, 7, and 8. This measure has several advantages. The most obvious advantage is that it takes into consideration the biological aspects, thereby providing a more biologically plausible foundation for deriving functional modules. Another advantage of this measure is that it does not treat a protein in isolation but rather as a part of a functional interplay between the protein and its neighbours, which is an important feature motivated by the

common definition of modules as units of interacting components that operate in an integrated manner towards achieving common functions.

4 **Development of a framework for identifying functional modules that uses the proposed measures**. In Chapter 6, we propose a framework for identifying modules, based on the weighted clustering coefficient. To show the potential of this framework, we perform the evaluation by comparing the obtained modules with the set of molecular complexes obtained from MIPS. In the same chapter, as well as in Chapter 7, the topological clustering coefficient (that only considers the triangle density of the modules) has been used to derive modules, and it resulted in fewer modules that matched complexes at high overlap score thresholds. Furthermore, in Chapter 9 we discuss the strengths and weaknesses of the combined methods for module identification proposed here, compared to the topology-based approaches that only consider the triangle density of the interactions.

5 **Combining several biological aspects results in identification of more biologically plausible modules than using each aspect separately**. Here, the results depend on whether we consider only direct neighbours when deriving modules, or if the whole neighbourhood with more distant neighbours is also considered (DFS-option). When using the option with direct neighbours, the combined aspect that considers ontology terms from all three sub-ontologies performs slightly better than each separate aspect. Using GO molecular function gives worse performance, no matter what option for node traversal is used. However, the best results in terms of matched MIPS complexes when using the DFS-option were generated by calculating weights based on the GO cellular component as a separate aspect, and the combination of all three aspects. This was demonstrated in Chapter 7.

6 **Considering the *k*-core sub-graph of the PIN in combination with the new measures (contribution 3) is more effective than basing those measures on the original PIN**. In Chapter 7, we demonstrated that restricting the analysis to the highest *k*-core PIN instead of the original PIN resulted in an improved set of modules, with respect to their overlap with known molecular complexes recorded in MIPS. This adds to the growing evidence that the *k*-core is a useful concept in general analysis of biological networks.

7 **The investigation of different features of so called multi-modular proteins, i.e., proteins that take part in multiple modules within the PIN, shows that these may be involved in the assembly and arrangement of cell structures (according to GO annotation) to a greater extent than single-modular proteins or proteins with lower numbers of occurrences across the generated module sets.** Also, the analysis of MIPS functional categories, presented in Chapter 8, along with the analysis of GO annotation, shows that the fraction of the proteins that belong to the category "cellular organisation" in multi-modular proteins is higher than the fraction of such proteins in the single-modular groups of proteins. Another frequently occurring GO term that is assigned to multi-modular proteins is "ribonucleoproteins complex biogenesis and assembly" which is a child term of "cellular component organisation and biogenesis". Hence, using the measures and techniques developed in this thesis, particularly contributions 3-6, we find evidence supporting the hypothesis that this GO term reveals the role of modules in building and supporting higher-order structure(s) of the PIN organisation. Other features that we have analysed to characterise possible differences between multi-modular and single-modular proteins are betweenness centrality and lethality. In both data sets, it is shown that there is significantelly higher fraction of lethal proteins among multi-modular proteins, also pointing at their significance. From the analysis of betweenness centrality, also performed in Chapter 8, it is also notable that proteins with high average module frequency have considerably high betweenness values, while the single-modular nodes exhibit a wide range of betweenness values in the yeast PIN. This also points to the greater importance of the multi-modular proteins, as those nodes may be potential bridges between modules in the network and have most influence on the information transfer between communicating modules. If a node with high betweenness centrality is removed, it may disconnect a different part of the network completely.

We here provide methods for identifying topologically and functionally cohesive modules in protein interaction network. Our results show that the identified modules include many examples of previously described functional modules. The network of modules provides a possibility to increase our understanding of the way modules interact and communicate. Modules do not act in isolation, but need to cooperate with

each other on the higher level of Life's complexity pyramid (see Figure 3 on page 17) to achieve certain functions.

An important feature of the identified modules is that they may contain overlapping proteins, i.e. one protein may take part in several processes or functions. Thus, to allow overlapping module sets to be found by the module identification algorithm is a biologically plausible approach. This allows us to link modules that share the same proteins. A possible application area that may utilise this is pathway discovery, but this property is also important for revealing higher order structure between modules. We show that many modules are enriched in proteins participating in a large number of non-redundant shortest paths between other partners. Such proteins are hereby assumed to act as "bridges" or "boundary spanners" between modules. Many of these proteins are known to have functional interconnectivity roles between modules.

Further analysis of the multi-modular proteins with respect to GO annotation reveals that the majority of the most frequently reappearing proteins significantly share the GO term "cell organization and biogenesis". To gain more confidence in the obtained results, we evaluated those proteins by analysing their MIPS functional categories (Mewes, et al., 2002), with the aim to determine what functional characteristics may be derived by studying proteins based on their module frequency. We observed that proteins involved in the cellular organisation category (O) appear more frequently among the top 100 multi-modular proteins, compared to the random sets of single-modular proteins. This result supports our findings based on studying GO biological process annotation, where "cell organization and biogenesis" was the most significant term among multi-modular proteins.

We employed a weighted clustering coefficient which considers all three edges of the triangles, because we suspected that this weighting scheme would improve our previous results, where only adjacent edges of a node were used in the calculation. However, the results clearly show that considering all three edges may not be a good idea. The weighted clustering coefficient defined by Onnela et al. (2005) resulted in lower values, and much fewer proteins that are used to potentially seed the modules, which may have affected results negatively.

Analysis of the identified functional modules allows the investigation of the biological processes that the modules participate in, as well as their functional roles. Since many of the modules are functionally homogeneous, containing majority of well-known

proteins with similar functions, along with some less characterised protein, this representation allows functional prediction for the less characterised proteins.

In conclusion, by using novel methods for module discovery, based on both topological and annotation-based semantic information, we are able to identify biologically relevant modules from protein interaction data, with the frequently overlapping proteins that are important building blocks in cellular organisation, i.e. support of cell structure and envelope.

10.2 Future Work

We have based our research on protein interaction networks from *S. Cerevisiae*, because it is a well-studied model organism for which large quantities of interaction data and annotations are available. A suitable continuation of this research would be to investigate functional modules in other organisms, such as *E. coli*. Another possible direction of future research may address the analysis of the modules in reactome data, with the purpose of comparing the modules derived from different biological levels, in this case reactome and interactome level.

Future work on the weighted metrics presented here may include researching different node weighting functions that, besides weighted (semantic) cohesiveness, take into account temporal characteristics of node and its interacting partners, such as correlation between mRNA expression levels. Another suitable test would be to develop a node scoring metric that is based on the combination of weighted cohesiveness and weighted average nearest-neighbours degree. One may evaluate how setting different thresholds on the latter metric, as a criterion for the inclusion of nodes into a module, would affect the resulting set of modules.

In this work, we integrate protein-protein interactions with the weight based on the GO terms, which may be interpreted as the probability of the interaction being true positive, or reliability. The semantic similarity between two proteins is defined as the average of pairwise term-term similarity values between all GO terms assigned to proteins. However, one may also test to calculate this value based on the maximum term-term similarity. Another alternative test of the proposed metric may be to only include the annotated terms with strongest evidence, i.e. the most reliable once, when calculating similarity.

A possible future application of the methods developed here is identification of modules of genes and proteins involved in various diseases, such as cancer. This module-level knowledge can contribute to the understanding of cancer on system-level, which may be useful for developing new drugs. Cancer-related networks for a specific type of cancer may be derived from, for example, gene expression data. Deriving gene networks makes it possible to apply network theoretic approaches on the interconnected genes that are potentially related to cancer development. Analysis of the modules and their interconnectedness provides the opportunity to reveal the role of modules in building higher-order structure(s) of the cancer-related network. Furthermore, a comparative analysis of the cancer-related networks derived from different types of cancer could be performed to identify common modules that are shared among different types, but also to identify the specific processes that characterize a certain type of cancer.

Modular analysis may also be applied to identify general properties of the interrelated genes that are involved in the origin of cancer cells. A suitable model for this analysis is a gene fusion network in human neoplasia (Hoglund, et al., 2006). By investigating topological properties of the cancer nodes in the network, such as node betweenness centrality, the cancer-related genes that act as "bridges" or communication points between various modules that correspond to cancer related processes may be identified.

Explaining the relationships between structure, function and regulation of molecular networks at different levels of the complexity pyramid of life is one of the main goals in systems biology. By integrating the topology, i.e. various structural properties of the networks with the functional knowledge encoded in protein annotations, and also analysing the interconnectivity between modules at different levels of the hierarchy, we aim to contribute to this goal. With the increasing availability of protein interaction data and more fine-grained GO annotations, our approaches will help constructing a more complete view of interconnected functional modules to better understand the organisation of cells.

10.3 Discussion and summary

Module definition

It is important to stress the difficulty of defining clear boundaries between modules, as which is due to the fact that they are nested structures where one module can be part of larger modules. As stated in (Tornow and Mewes, 2003), no clearly separated networks exist in the cell. Even if we partition the network into different sets of modules, there

are no objective criteria to determine that one partition is better than other. Hence, this makes the evaluation of the predicted modules very difficult. In our evaluation, we use the MIPS protein catalogue, where complexes are obtained from the lowest and most granular level of hierarchy. For example, the assembly complex termed "replication complexes" involves 55 proteins in total and is subdivided into seven sub-groups of complexes: post-replication complex, replication initiation complex, pre-replication complex, GINS complex, replication complex, replication fork complexes, and telomerase. A sub-group named "Replication fork complexes" contains, in turn, 14 different complexes, and so on.

However, we think that it is more advantageous and biologically realistic to provide the possibility of generating different partitions of modules depending on the degree of similarity between the proteins rather than to represent the biological reality by a single partition. By tweaking the parameters, we can obtain a spectrum of different modular representations, from very specific to very general module patterns. We can say that there are different levels of network resolution, where individual protein sub-complexes may be considered as modular units at a high level of resolution, while assembly of several sub-complexes that contribute to the common function may correspond to a low-level resolution map of modules. Yeast mediator complex (YMC) may be a good example that illustrates such differences between different resolution levels. YMC is a large, modular protein complex that transmits regulatory signals from the transcriptional activators to the RNA polymerase II initiation machinery (Guglielmi, et al., 2004). In *S. cerevisiae*, YMC is thought to be composed of 24 subunits organised in four sub-complexes, termed the head, middle, tail and Cdk8 (Srb8-11) modules. 16 of these 24 subunits are present in our CORE data set. All of them could be found in one module when the *NWP* parameter was set to a low value, while they were split into several modules when a high percentage of similarity was required for inclusion in modules ($PWD > 0.5$), which clearly shows the different resolutions of modular units. Even though no clear distinction between different modules in YMC could be found, probably due to missing data, some modules, such as Cdk8, become part of a separate complex after increasing the value of the *NWP* parameter.

Module identification method, parameters, and evaluation

In Chapter 4, we have described the evaluation of the three most common semantic similarity measures, and this evaluation constitutes the basis for our choice of measure.

After choosing the semantic similarity measure, we considered different ways of weighting edges with respect to that measure. In the study by Lubovac et al. (2005), we showed that "node-to-neighbour" similarity, i.e. similarity of edges adjacent to a node contributes more to the overall similarity than the similarity between the neighbours of the node. We also showed in section 7.3 that the weighted function adapted from (Barrat, et al., 2004) resulted in a larger number of predicted complexes that matched MIPS complexes, than when using the weighted function by Onnela et al. (2005). The main difference between the two functions is that the latter one also considers the third edge of the triangle formed in the neighbourhood of node *i*, which did not seem to be a good idea.

To further evaluate the generated complexes, and maximize the biological relevance of predicted complexes according to the given benchmark (the MIPS database), we used overlap measures (Bader and Hogue, 2003; Poyatos and Hurst, 2004). We used two different thresholds. The results from only one, defined in Equation 17 on page 90, are presented, because the results obtained from the two measures were highly correlated (the Pearson correlation coefficient varied between 0.91 and 0.95). As the density of interactions is regarded as an important aspect, it was used in combination with the overlap measure to choose the best parameter setting. The best agreement between process-based and function-based modules in terms of highest overlap was also obtained when using the chosen parameter setting.

NWP similarity of weighted cohesiveness between nodes is one of the major parameters that provide the possibility of generating different partitions of modules depending on the degree of functional similarity between the proteins. This is a more biologically plausible scenario than generating a single partition. By tweaking this parameter, we can obtain different levels of network resolution, where individual protein sub-complexes may be considered as modular units at a high level of resolution, while assembly of several sub-complexes that contribute to the common function may correspond to a low level resolution map of modules. The yeast mediator complex (YMC), which has been described in detail above, may be a good example that illustrates such differences between different resolution levels.

Introducing the overlap parameter *f* have resulted in an increased number of matches against MIPS complexes. Also, the *NWP* threshold that resulted in the largest number

of modules was $NWP > 0.95$, when combining with the f parameter, compared to $NWP > 0.6$ when the overlap threshold was not used.

To ensure that SWEMODE is not improperly affected by the expected high false-positive rate in large-scale interaction data sets, we applied the method on the network derived from the literature.

Another way of finding an optimal parameter setting is described in (Jansen, et al., 2002), as finding a trade-off between the highest possible coverage and the lowest possible error rate. To be able to assess the prediction performance of the results, we need control datasets including gold standard positives (i.e. proteins that are connected) and gold standard negatives (i.e. proteins that are not connected). In this work, we have used the MIPS complex catalogue as gold standard to evaluate the predicted modules. Assuming this standard to determine whether two proteins belong to the same complex or not, different thresholds for predicting whether two proteins interact or not based on the datasets mentioned above could be tested, to optimise the performance of our module identifying framework.

The proposed approach could be further validated by using only the GO annotation that is associated with the strongest evidence codes

Summary

In summary, we propose knowledge-based methods that integrate domain specific knowledge with topological information to derive modular structures from PINs. In contrast to other approaches that first derive modules, and then analyse their biological plausibility, we take into consideration the functional knowledge about the experimental interactions, and in this way strengthen the validity of the obtained structures. Modules obtained in this way serve as models for studying interconnectivity, which is a step towards reconstruction of the higher order hierarchy of cellular networks.

We have employed three different biological aspects – molecular function, biological process and cellular component, and tested their suitability for deriving modules. Based on the evaluation of the overlap with the MIPS database, we found that biological process and cellular component annotation is more advantageous to module prediction than molecular function. The best overlap between predicted modules and the MIPS data is obtained by combining all three aspects.

It was indicated in previous work that identification of protein complexes (an example type of module) may become more challenging as additional protein-protein interaction data becomes available, because the interactions are noisy, and the integration of protein-protein interaction data with annotation might prove a useful solution to this problem. Our integrated approach contributes to this solution, by increasing the confidence in high-throughput Y2H data. It also provides means for an increased understanding of the higher-order structures underlying cellular function. As annotations become more complete, the increased biological relevance of our module predictions with integrated approaches is expected to be even more evident.

Finally, we would like to discuss one of the biggest issues that we came across during this work, namely the difficulty to clearly characterise modules. Therefore, we would like to stress that modularity is, in consistency to other important notions in molecular biology (like homology for example), although intuitively very easy to understand, conceptually very difficult to characterise. As we already pointed out, there is no generally accepted definition of modules. A pioneering work in this area, performed by Hartwell et al. (1999) provides a wide definition, which opens space for different authors to define different more specific criteria. This is, as also pointed out in (Schlosser and Wagner, 2004), unavoidable, and "retaining a pragmatic pluralism of different modularity concepts is probably a fruitful strategy for broadening our perspective and illuminating the importance of modularity at many different levels of organization" (Schlosser and Wagner, 2004).

Appendix A

Filamentation network

A.1 Protein interaction network consisting of filamentation proteins and their neighbours

This appendix contains the input graph for yeast filamentation network.

Node:CDC28 Neighbours:CKS1.CAK1.CDC6.CLB1.CLB2.CLB3.CLB4.CLN1.FAR1.CLN2.MPT5.SCS2.CLN3.PCL6.PCL7.STB1.CLA4.GSY1.RPN1.TCP1.CDH1.DAL7.SLD2.PHO2.PDS1.GRR1.CLB5.SWE1.SIC1.PAK1.FUN30.VMR1.CAF120.DNA2.NUP1.KEL2.SFB3.PKC1.SIP3.CHS2.DBF4.SWI5.YRF13.LTE1.SKG3.SEC3.FKH2.TOS2.YTA7.ORC6.GIN4.MOB1.RAD9.BEM3.MYO3.RGA1.SEN1.NBP1.SWI6.KIN2.BCK2.YNR047W.YDL025C.HOS4.HCM1.SPA2.YJL084C.CDC20.KAR3.ORC1.DBF2.BNA3.DBF20.IFH1.CDC5.ZIP1.MCM3.STC1.KIP2.SSK1.PKH2.IPL1.ACE2.NTH1.DPB2.SRL3.PRR1.RLF2.KIP3.ELM1.KEL1.BUD6.YHL050C.JSN1.SPC110

Node:RVS167 Neighbours:ABP1.P02579.RVS161.SRV2.GTS1.LAS17.YBR108W.HUA1.ACF4.APP2YNL086W.APP1.CNA1.MTH1.YJR115W.PDR3.YBP2.ESP1.EXO70.PSE1.KAP122.LCP5.MSU1.MUS81.SEC21.SEC8.SLF1.SLN1.SRP54.URA7.GFD2.RPL36B.SLA1.YSC84.PCL2.MYO3.MYO5.VRP1.RSP5.IDH1.LYS12.UBP7

Node:KSS1 Neighbours:STE5.STE11.STE7.STE12.MPT5.ARG81.ARP7.BEM3.CCT2.FET4.GFA1HAS1.HXT6.MKT1.MSE1.PHO84.PMA1.RPA135.SEN1.GSN1.YHR033W.YJR072C.BCK2.SST2

Node:SNF1 Neighbours:SIP3.SIP1.SNF4.SIP2.SIP4.GAL83.MGS1.MIG1.STD1.SSN3.SSN8.SIN4NRG2.NRG1.RCK1.ARF1.PRB1.YPK2.PAK1.REG1.GLN3.HHT1

Node:ABP1 Neighbours:RVS167.SRV2.SAC6.SLA1.SLA2.CLA4.APP1.ARP2.P02579.PRK1.ARK1INP52.ARP3.SPS1.TUP1.YSC84.MYO5.YIR003W

Node:CDC42 Neighbours:BNI1.BEM4.RDI1.BOI1.CDC24.CLA4.FAR1.GIC1.IQG1.RGA1.STE20BEM1.SKM1.SEC3.ADH2

Node:STE4 Neighbours:STE18.GPA1.STE5.AKR1.FAR1.CDC24.SYG1.PLP1.CCT2.CCT3.CCT5CCT6.PGM2.TCP1

Node:BEM1 Neighbours:P02579.STE20.STE5.LAS17.BOI1.BOI2.CDC24.FAR1.RSR1.SEC15.SWE-1.YEL043W.CAF130.CDC42

Node:BMH2 Neighbours:STE20.BMH1.NTH1.NTH2.PIK1.RTG2.PSK1.MSB3.GCR2.REG2.PPT1LRE1.RPS0A

Node:STE11 Neighbours:STE5.KSS1.FUS3.STE50.GPA1.HOG1.HSP82.BUD6.HSC82.SPA2.PBS2LSM1.VPS21

Node:SNF4 Neighbours:SNF1.SIP2.GAL83.YMR291W.P25575.SIP1.RCK1.HRK1.TOS3.BMH1.C-KA1.PAK1.REG1

Node:TPK3 Neighbours:BCY1.PKH1.TPK1.TPK2.CPR6.GPH1.PFK1.SEC27.YHR033W.JJJ1.YPT7YCK1.YPT53

Node:BNI1 Neighbours:CDC42.P02579.BUD6.PFY1.RHO1.BCK1.SPA2.HOF1.TEF1.FUS1.MYO3ARP2

Node:CLA4 Neighbours:CDC42.ABP1.BEM3.BOI2.CDC12.NUP100.GIC1.MSB2.RGA1.SLA2.ZD-S2.CDC28

Node:STE20 Neighbours:PRP21.BEM1.P02579.CDC42.CLN2.HSL7.BMH1.BMH2.BOI1.BOI2.CD-C24

Node:DHH1 Neighbours:CDC7.LSM2.POP2.PAT1.SER3.SEC27.YKU80.LSM1.LSM7.RPC40.SEC7

Node:GIC1 Neighbours:BUB2.CDC42.BEM4.CDC12.CLA4.DFG5.STE50.CSM1.RRP14.ZDS2.ZD-S1

Node:RSP52 Neighbours:BUL2.BUL1.RPO21.SPT23.HXT6.RVS167.YFR022W.ROD1.P02579.SLA-

Node:HTB2 Neighbours:INO4.RAD16.ISW1.ISW2.KAP114.MOT1.NAP1.STH1.VPS1.PUF6

Node:SPA2 Neighbours:STE7.MKK1.MKK2.STE11.BNI1.BUD6.PEA2.SLT2.MSB3.CDC28

Node:CLN2 Neighbours:CDC28.FAR1.CDC53.GRR1.STE20.SWI4.SWI6.BUD2.CKS1
Node:FUS3 Neighbours:STE5.STE11.STE7.PTP3.MPT5.YHR168W.MSG5.RCK1.GPA1
Node:TUP1 Neighbours:CYC8.HMLALPHA2.HHF1.HHT1.HDA1.SKO1.ABP1.CUT1.ECM10
Node:CLB2 Neighbours:SWI4.CKS1.CDC28.NAP1.CDC20.CDH1.SWE1.CDC23.RPL40A
Node:CDC6 Neighbours:CDC28.CDC4.ORC1.ORC2.ORC3.ORC4.ORC5.ORC6.RPT1
Node:CDC25 Neighbours:HSP82.RAS1.RAS2.SDC25.SSA1.SSA2.SSA3.CYR1
Node:BEM4 Neighbours:CDC42.RHO1.RHO4.CDC11.CDC12.GIC1.RSR1.RHO2
Node:AKR1 Neighbours:IQG1.GCS1.STE18.STE3.STE4.STE5.YCK2.YCK1
Node:HOG1 Neighbours:RCK1.HOT1.PBS2.STE11.PTP2.PTP3.SKO1.RCK2
Node:PBS2 Neighbours:HOG1.STE11.SHO1.FET4.NBP2.PTC1.SSK2.SSK22
Node:SLA2 Neighbours:P02579.ABP1.CLA4.APP1.YSC84.RPG1.LAS17.RSP5
Node:CYR1 Neighbours:IRA1.CDC25.RAS2.SRV2.DBF2.PHO81.MAK5
Node:TPK1 Neighbours:BCY1.FET4.RIM15.TCP1.TPK2.TPK3.YPT53
Node:BCY1 Neighbours:TPK1.TPK2.TPK3.GLC7.SRO77.CKA1.YCK1
Node:SRV2 Neighbours:ABP1.P02579.CYR1.RVS167.AIP1.PFY1.TRM5
Node:BUD6 Neighbours:P02579.BNI1.YGL015C.STE11.SPA2.BNR1.CDC28
Node:HMLALPHA1 Neighbours:HMLALPHA2.MCM1.STE12.SEC61.ERV29
Node:HMLALPHA2 Neighbours:TUP1.MCM1.HMLALPHA1.HMRA1.CYC8
Node:STE50 Neighbours:STE11.SPC24.STE5.GIC1.AKL1
Node:SWI4 Neighbours:SWI6.CLB2.CLN2.SLT2.RAD53
Node:CLN1 Neighbours:CDC28.FAR1.PHO85.CKS1.GRR1
Node:HSL1 Neighbours:CDC3.HSL7.CDC20.CDH1.KRE33
Node:TPK2 Neighbours:BCY1.TPK1.TPK3.YCK1
Node:MPT5 Neighbours:CDC28.KSS1.FUS3.SST2
Node:SHO1 Neighbours:PBS2.KIN2.PKC1.SST2
Node:STE7 Neighbours:STE5.KSS1.FUS3.SPA2
Node:BMH1 Neighbours:STE20.NTH1.BMH2.SNF4
Node:SFL1 Neighbours:SSN2.SSN8.SIN4.ROX3
Node:SIP4 Neighbours:SNF1.GAL83.SSN3.CAT8
Node:STE12 Neighbours:KSS1.MCM1.HMLALPHA1
Node:SAC6 Neighbours:LAS17.P02579.ABP1
Node:CLB1 Neighbours:CKS1.CDC28.SPC29
Node:CAK1 Neighbours:SMK1.CDC28.SGV1
Node:CLN3 Neighbours:YDJ1.FAR1.CDC28
Node:RAS1 Neighbours:MEK1.CDC25.RAS2
Node:RAS2 Neighbours:CDC25.CYR1.RAS1
Node:RSR1 Neighbours:BEM1.CDC24.BEM4
Node:IRA1 Neighbours:CYR1.RIM11
Node:SSA4 Neighbours:CEG1.CKB2
Node:IRA2 Neighbours:RIM11
Node:SDC25 Neighbours:CDC25

A.2 Protein graph consisting of filamentation proteins

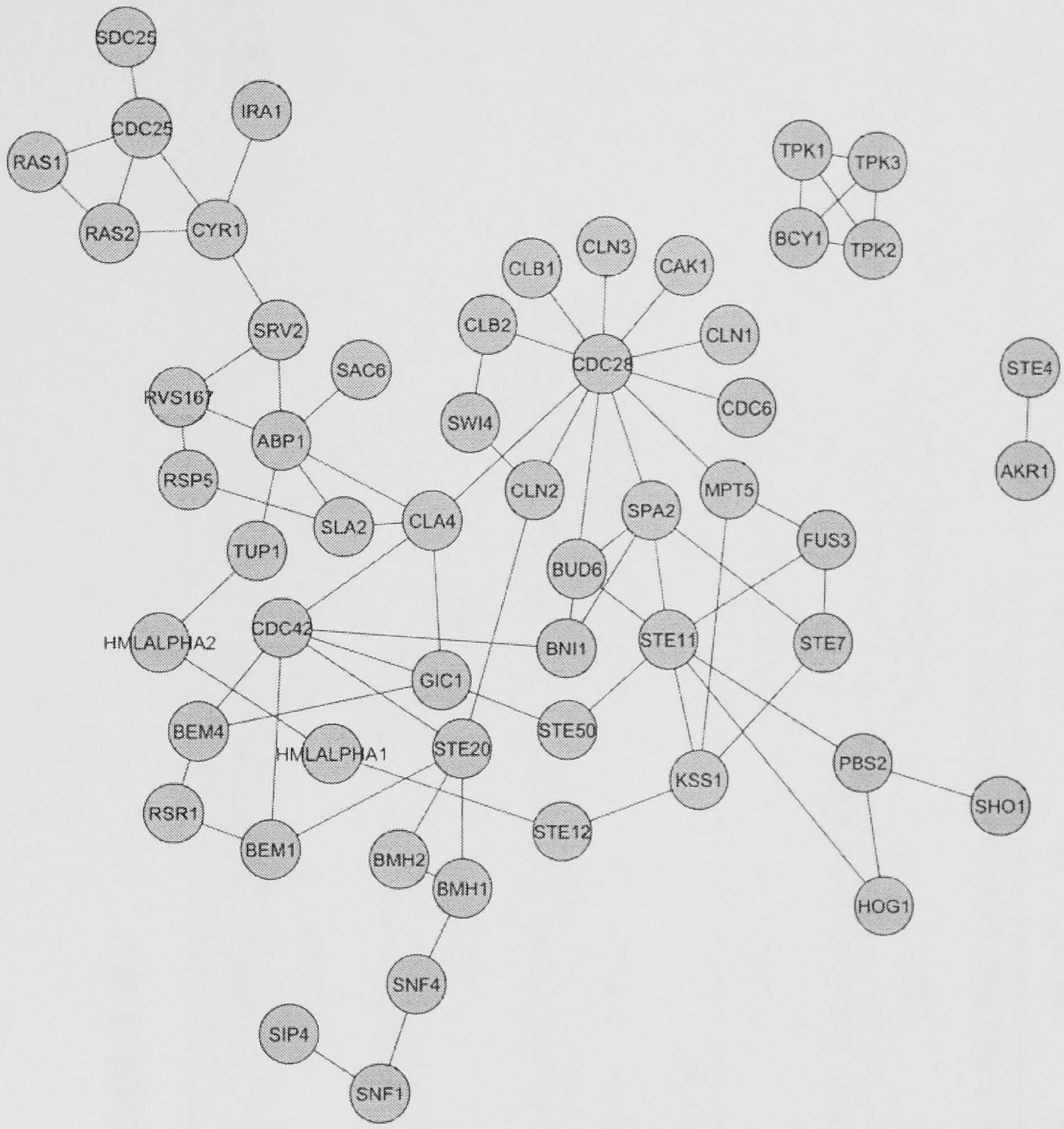


Figure 42: Modular network involving modules in filamentation network

Appendix B

Complete set of modules (Chapter 6)

Rank	Score	Proteins	Molecular function	*P* value	Frequency (%)	Cellular component	*P* value	Frequency (%)
1	9.67	13	RNA binding	$8.13 \cdot 10^{-15}$	92	Small nuclear ribonucleo-protein complex	$5.72 \cdot 10^{-21}$	90
		Lsm5, Lsm6, Lsm2, Lsm4, Lsm1, Lsm8, Lsm3, Lsm7, Pat1, Prp4, Prp24, Smd2, Smd3						
2	7.88	18	Endopeptidase activity	$5.06 \cdot 10^{-19}$	61	Proteasome complex	$1.10 \cdot 10^{-30}$	83
		Rpt1, Rpt2, Rpt3, Rpt4, Rpt5, Rpt6, Rpn1, Rpn3, Rpn10, Rpn11, Rpn12, Rpn14, Cdc6, Pre1, Nas6, Leo1, Ctr9, Rad23						
3	7.71	8	Oligosaccharyl transferase activity	$5.39 \cdot 10^{-24}$	100	Oligosaccharyl transferase complex	$5.39 \cdot 10^{-24}$	100
		Wbp1, Ost1, Ost2, Ost3, Ost4, Ost5, Swp1, Stt3						
4	7.64	12	RNA binding	$7.78 \cdot 10^{-10}$	75	mRNA cleavage factor complex	$1.81 \cdot 10^{-31}$	100
		Cft2, Pap1, Pfs2, Fip1, Pta1, Pti1, Mpe1, Pcf11, Ref2, Rna14, Swd2, Ssu72						
5	7.54	14	3'-5'-exoribonuclease activity	$1.88 \cdot 10^{-15}$	50	Transcription factor complex	$1.34 \cdot 10^{-25}$	100
		Cdc39, Cdc36, Not3, Not5, Mot2, Caf4, Ssn3, Ssn2, Pop2, Ccr4, Caf40, Caf130, Taf1, Taf6						
6	7.00	13	ATP-dependent RNA helicase	$0.96 \cdot 10^{-03}$	15	Nucleolus	$2.83 \cdot 10^{-16}$	92

			activity					
		Nop15, Erb1, Mak21, Brx1, Dbp10, Has1, Nop4, Nop7, Rlp7, Sda1, Nug1, Cic1, Ytm1						
7	6.86	8	Structural constituent of cytoskeleton	$7.54 \cdot 10^{-10}$	62	Septin ring	$2.11 \cdot 10^{-15}$	75
		Cdc11, Cdc3, Cdc10, Cdc12, Shs1, Kcc5, Gin4, Bni5						
8	6.86	8	NAD-independent histone deacetylase activity	$8.68 \cdot 10^{-18}$	75	Histone deacetylase complex	$9.32 \cdot 10^{-18}$	88
		Hos2, Snt1, Hos4, Set3, Hst1, Zds1, Sif2, Cpr1						
9	6.57	8	Translation initiation factor activity	$6.26 \cdot 10^{-20}$	100	Eukaryotic translation initiation factor 2B complex	$8.47 \cdot 10^{-15}$	62
		Gcd1, Gcn3, Gcd2, Gcd6, Gcd7, Gcd11, Sui2, Sui3						
10	6.44	10	Structural molecule activity	$2.84 \cdot 10^{-10}$	89	Pore complex	$5.12 \cdot 10^{-17}$	89
		Nup84, Nup145, Nup85, Nup42, Nup100, Nup120, Msn5, Seh1, Sec13, Nup57						
11	6.33	7	DNA clamp loader activity	$2.96 \cdot 10^{-11}$	57	DNA replication factor C complex	$9.13 \cdot 10^{-21}$	100
		Rfc3, Rfc4, Rfc2, Rfc5, Cft8, Cft18, Elg1						
12	6.00	7	Hydrogen-transporti-ng ATP synthase act-ivity, rotational	$5.29 \cdot 10^{-16}$	86	Proton-transporting ATP synthase complex	$1.16 \cdot 10^{-18}$	100

			mechanism					
		Atp18, Atp6, Atp1, Atp2, Atp7, Atp17, Atp20						
13	6.00	6	DNA replication origin binding	$6.65 \cdot 10^{-18}$	100	Origin recognition complex	$3.10 \cdot 10^{-19}$	100
		Orc6, Orc5, Orc1, Orc2, Orc3, Orc4						
14	6.00	6	Ubiquitin-protein ligase activity	$1.31 \cdot 10^{-11}$	83	Anaphase-promoting complex	$1.11 \cdot 10^{-16}$	100
		Apc1, Doc1, Cdc16, Cdc23, Cdc26, Cdc27						
15	5.73	16	snoRNA binding	$1.04 \cdot 10^{-18}$	56	Small nucleolar ribo-nucleoprotein complex	$1.13 \cdot 10^{-22}$	75
		Utp22, Utp4, Utp6, Utp8, Utp7, Utp10, Utp18, Utp21, Pwp2, Prp45, Dip2, Ecm16, Emg1, Kre33, Rok1, Enp2						
16	5.71	8	DNA packaging	$1.78 \cdot 10^{-10}$	87.5	Nuclear chromatin	$1.68 \cdot 10^{-15}$	87.5
		Hht1, Sir3, Sir4, Hhf1, Arp4, Hta1, Htb1, Cka1						
17	5.60	6	Structural molecule activity	$2.09 \cdot 10^{-03}$	50	Pore complex	$3.55 \cdot 10^{-08}$	67
		Nup60, Nup2, Nup116, Kap95, Srp1, Gsp1						
18	5.60	6	Transcription regulator activity	$9.86 \cdot 10^{-07}$	83	SLIK complex	$2.20 \cdot 10^{-13}$	83
		Taf9, Gcn4, Ngg1, Ada2, Taf5, Spt7						
19	5.50	9	Structural molecule activity	$2.22 \cdot 10^{-08}$	78	Arp2/3 protein complex	$2.70 \cdot 10^{-20}$	78

		Arp3, Arp2, Arc15, Arc18, Arc19, Arc35, Arc40, Las17, Abp1						
20	5.50	9	General RNA polymerase II transcription regulator activity	$1.14 \cdot 10^{-13}$	78	DNA-directed RNA polymerase II, holoenzyme	$4.72 \cdot 10^{-13}$	78
		Srb2, Srb4, Srb5, Srb6, Med8, Gal4, Gal11, Ssn8, Paf1						
21	5.43	8	Hydrogen-transporting ATPase activity, rotational mechanism	$1.02 \cdot 10^{-22}$	100	Hydrogen-translocating V-type ATPase complex	$3.20 \cdot 10^{-13}$	100
		Vma7, Vma4, Vma8, Vma10, Vma13, Tfp1, Vph1, Stv1						
22	5.27	12	RNA bidning	$2.33 \cdot 10^{-12}$	91	Small nuclear ribonucleoprotein complex	$8.36 \cdot 10^{-20}$	91
		Snu71, Prp39, Prp40, Mud1, Nam8, Smx2, Smx3, Snp1, Vps41, Yhc1, Snu56, Smd1						
23	5.14	8	Intracellular transporter activity	$1.35 \cdot 10^{-13}$	75	Vacuole	$8.95 \cdot 10^{-09}$	75
		Vam3, Vam7, Ykt6, Vti1, Nyv1, Vps33, Pep3, Sec17						
24	5.00	13	Translation initiation factor activity	$2.64 \cdot 10^{-14}$	54	Eukaryotic 43S preinitiation complex	$1.76 \cdot 10^{-09}$	46
		Rpg1, Pci8, Prt1, Sla2, Tif5, Tif6, Tif34, Tif35, Hcr1, Sui1, Ckb2, Hat1, Sua7						
25	5.00	11	Endopeptidase activity	$1.02 \cdot 10^{-12}$	64	Proteasome core complex	$8.02 \cdot 10^{-17}$	64
		Pre8, Nup49, Pre4, Pre5, Pre6, Pre9, Pph22, Ecm29, Pup3, Scl1, Pse1						

26	5.00	5	No significant terms			ER to Golgi transport vesicle	$5.75 \cdot 10^{-13}$	100
		Erv25, Emp24, Sec23, Sec24, Sec31						
27	5.00	5	Histone lysine N-methyltransferase activity (H3-K4 specific)	$1.59 \cdot 10^{-15}$	100	Histone methyltransferase complex	$1.59 \cdot 10^{-15}$	100
		Set1, Bre2, Swd1, Swd3, Shg1						
28	5.00	5	Peroxisome targeting sequence binding	$1.69 \cdot 10^{-06}$	40	Peroxisome	$1.67 \cdot 10^{-11}$	100
		Pex17, Pex5, Pex7, Pex13, Pex14						
29	5.00	5	Transcription regulator activity	$1.73 \cdot 10^{-07}$	100	SWI/SNF complex	$4.44 \cdot 10^{-08}$	60
		Hir2, Snf2, Snf5, Swi3, Hir1						
30	5.00	5	DNA-directed RNA polymerase activity	$1.89 \cdot 10^{-12}$	100	DNA-directed RNA polymerase III complex	$6.89 \cdot 10^{-14}$	100
		Rpc53, Rpc34, Rpc37, Rpc40, Rpo31						
31	4.80	6	RNA polymerase II transcription elongation factor activity	$5.49 \cdot 10^{-13}$	83	Transcription elongation factor complex	$1.18 \cdot 10^{-12}$	83
		Iki1, Elp2, Elp4, Elp6, Iki3, Kti12						

32	4.50	5	Structural constituent of cytoskeleton	$1.37 \cdot 10^{-11}$	100	Gamma-tubulin complex	$6.96 \cdot 10^{-10}$	60
		Spc98, Spc72, Spc97, Spc110, Tub4						
33	4.40	6	DNA-directed RNA polymerase activity	$8.59 \cdot 10^{-15}$	100	DNA-directed RNA polymerase II, core complex	$3.21 \cdot 10^{-17}$	100
		Rpb4, Rpb2, Rpb3, Rpb5, Rpb7, Rpo21						
34	4.33	7	Signal transducer activity	$2.09 \cdot 10^{-05}$	43	Site of polarized growth	$1.29 \cdot 10^{-12}$	100
		Cdc24, Bem1, Cdc42, Far1, Rsr1, Swe1, Ste20						
35	4.00	6	Protein binding	$5.41 \cdot 10^{-03}$	50	Cellular component unknown	$4.60 \cdot 10^{-03}$	83
		Snz3, Sno1, Sno2, Sno3, Snz1, Snz2						
36	4.00	4	Rab GTPase binding	$4.51 \cdot 10^{-07}$	50	Membrane	$4.20 \cdot 10^{-04}$	100
		Yip4, Yip5, Ypt1, Sec4						
37	4.00	4	3'-5'-exoribonuclease activity	$3.71 \cdot 10^{-11}$	100	Cytoplasmic exosome (RNase complex)	$3.53 \cdot 10^{-12}$	100
		Rrp43, Rrp4, Dis3, Ski6						
38	4.00	4	Protein transporter activity	$2.21 \cdot 10^{-07}$	100	Mitochondrial inner membrane protein insertion complex	$6.60 \cdot 10^{-10}$	75

		Tim22, Tim54, Mrs5, Mrs11							
39	4.00	4	Alpha-1,6-mannosyltransferase activity	$4.58 \cdot 10^{-13}$	100	Mannosyltransferase complex	$4.58 \cdot 10^{-13}$	100	
		Hoc1, Mnn10, Mnn11, Anp1							
40	4.00	4	tRNA-intron endonuclease activity	$9.05 \cdot 10^{-14}$	100	tRNA-intron endonuclease complex	$9.05 \cdot 10^{-14}$	100	
		Sen34, Sen2, Sen15, Sen54							
41	4.00	4	DNA binding	$7.01 \cdot 10^{-07}$	100	Nucleotide-excision repair factor 1 complex	$2.78 \cdot 10^{-10}$	75	
		Rad10, Msh2, Rad1, Rad14							
42	3.80	11	Nucleoside-triphosphatase activity	$7.34 \cdot 10^{-06}$	50	Chromatin remodeling complex	$1.07 \cdot 10^{-12}$	70	
		Isw1, Isw2, Chd1, Itc1, Mot1, Npl6, Rsc2, Rsc6, Vps1, Rsc58, P89501							
43	3.75	9	Unfolded protein binding	$4.88 \cdot 10^{-07}$	44	Cytoplasm	$1.75 \cdot 10^{-03}$	100	
		Hsc82, Ppg1, Cpr6, Sti1, Hsp82, Ste11, Sba1, Cns1, Myo4							
44	3.71	8	DNA-directed RNA polymerase activity	$1.76 \cdot 10^{-10}$	100	DNA-directed RNA polymerase I complex	$1.84 \cdot 10^{-22}$	100	

		Rpa190, Rpa12, Rpa43, Rpa49, Rpa135, Rpb8, Rpb10, Rpo26						
45	3.67	7	RNA binding	0.052	29	Small nucleolar ribonucleoprotein complex	$1.37 \cdot 10^{-07}$	57
		Cbf5, Naf1, Sik1, Nop1, Puf6, Rrp12, Gar1						
46	3.60	6	Molecular function unknown	$5.87 \cdot 10^{-03}$	100	TRAPP complex	$6.65 \cdot 10^{-18}$	100
		Bet3, Trs20, Trs120, Trs130, Bet5, Kre11						
47	3.50	5	GTPase activity	$4.47 \cdot 10^{-06}$	60	Late endosome	$7.05 \cdot 10^{-08}$	60
		Ypt53, Ypt52, Vps21, Tpk3, Yif1						
48	3.50	5	DNA binding	$3.42 \cdot 10^{-06}$	80	DNA replication factor A complex	$1.69 \cdot 10^{-06}$	40
		Rfa1, Rfa2, Rad52, Msh6, Mph1						
49	3.33	4	Cytoskeletal protein binding	$4.33 \cdot 10^{-05}$	100	Actin cortical patch	$2.86 \cdot 10^{-05}$	100
		Vrp1, P02579, Rvs167, Sla1						
50	3.33	4	Binding	$7.00 \cdot 10^{-04}$	100	Extrinsic to internal side of plasma membrane	$1.80 \cdot 10^{-06}$	50
		Ure2, Tor1, Tor2, Gln3						
51	3.33	4	General RNA polymerase II transcription factor activity	$2.44 \cdot 10^{-06}$	75	Transcription factor TFIIF complex	$2.78 \cdot 10^{-10}$	75

		Rpb9, Taf14, Tfg1, Tfg2						
52	3.33	4	Unfolded protein binding	$3.70 \cdot 10^{-04}$	50	Ribosome	$8.69 \cdot 10^{-03}$	50
		Zuo1, Bud14, Asc1, Ssz1						
53	3.33	4	DNA secondary structure binding	$4.51 \cdot 10^{-07}$	50	Cohesin complex	$4.05 \cdot 10^{-06}$	50
		Smc2, Smc1, Smc3, Kar5						
54	3.33	4	Transcription regulator activity	$3.30 \cdot 10^{-04}$	75	Nucleus	$5.47 \cdot 10^{-03}$	100
		Arg80, Mcm1, Arg81, Arg82						
55	3.33	4	Binding	$6.20 \cdot 10^{-04}$	100	COMA complex	$6.60 \cdot 10^{-10}$	75
		Okp1, Ctf19, Mcm21, Cep3						
56	3.33	4	Molecular function unknown	0.033	100	Retromer complex	2.2110^{-13}	100
		Pep8, Vps17, Vps29, Vps35						
57	3.33	4	Protein transporter activity	$2.21 \cdot 10^{-09}$	100	Mitochondrial outer membrane translocase complex	$3.53 \cdot 10^{-09}$	75
		Tom40, Tom20, Tom22, Tim17						
58	3.33	4	Molecular function unknown	0.033	100	AP-2 adaptor complex	$9.05 \cdot 10^{-14}$	100

		Aps2, Apm4, Apl1, Apl3						
59	3.33	4	Clathrin binding	$5.18 \cdot 10^{-12}$	100	AP-1 adaptor complex	$2.21 \cdot 10^{-13}$	100
		Aps1, Apl2, Apm1, Apm2						
60	3.20	6	Ubiquitin-protein ligase activity	$7.04 \cdot 10^{-09}$	67	SCF ubiquitin ligase complex	$2.90 \cdot 10^{-14}$	83
		Cdc53, Cdc4, Cdc34, Cln2, Grr1, Ufo1						
61	3.20	6	ATP-dependent DNA helicase activity	$3.46 \cdot 10^{-10}$	67	Pre-replicative complex	$2.20 \cdot 10^{-13}$	83
		Mcm3, Mcm2, Cdc28, Cdc45, Cdc46, Cdc47						
62	2.80	6	SNAP receptor activity	$7.08 \cdot 10^{-07}$	50	Golgi apparatus	$3.49 \cdot 10^{-08}$	83
		Sec22, Sed5, Sec20, Arf1, Cog3, Sly1						
63	2.67	4	Structural molecule activity	$4.50 \cdot 10^{-04}$	75	Pore complex	$1.36 \cdot 10^{-06}$	75
		Nic96, Tfs1, Nup188, Pom152						
64	2.67	4	Microfilament motor activity	$2.22 \cdot 10^{-09}$	75	Actin cytoskeleton	$1.84 \cdot 10^{-08}$	100
		Mlc1, Myo1, Myo2, She3						
65	2.67	4	GTPase activator activity	$1.00 \cdot 10^{-04}$	50	Spindle pole	$6.31 \cdot 10^{-09}$	100

		Bfa1, Cdc5, Cdc14, Bub2						
66	2.67	4	Transcription regulator activity	0.011	50	Bud	$2.09 \cdot 10^{-03}$	50
		Swi4, Swi6, Clb2, Slt2						
67	2.67	4	General transcriptional repressor activity	$4.51 \cdot 10^{-07}$	50	Nucleus	$5.47 \cdot 10^{-03}$	100
		Tup1, Cyc8, Hda1, Sko1						
68	2.50	5	No significant terms			Protein complex	0.054	60
		Tcp1, Cdc55, Tpk1, Yju2, Cct2						
69	2.50	5	Hydrolase activity	0.088	40	Protein kinase CK2 complex	$3.00 \cdot 10^{-06}$	40
		Cka2, Ckb1, Abf1, Sap185, Sin3						
70	2.50	5	Cytoskeletal regulatory protein binding	$3.22 \cdot 10^{-05}$	60	Polarisome	$3.22 \cdot 10^{-05}$	60
		Bni1, Bud6, Bck1, Spa2, Fus1						
71	2.40	6	Structural molecule activity	0.032	33	Pore complex	$3.55 \cdot 10^{-08}$	67
		Nup159, Crm1, Gle1, Dbp5, Mtr10, Nup82						
72	2.26	9	Histone deacetylase activity	$1.05 \cdot 10^{-06}$	33	Histone deacetylase complex	$7.05 \cdot 10^{-09}$	44

		Rpd3, Cpr7, Sap30, Eaf3, Nrd1, Pho23, Sds3, Sto1, Rxt2						
73	2.00	4	No significant terms			Nucleolus	$2.82 \cdot 10^{-06}$	100
		Mak5, Nop2, Nop12, Nsa1						
74	2.00	4	No significant terms			Nuclear envelope	$1.05 \cdot 10^{-05}$	75
		Ntf2, Yrb2, Nsp1, Gsp2						
75	2.00	4	Transcription regulator activity	$3.89 \cdot 10^{-06}$	100	Transcription factor complex	$1.85 \cdot 10^{-05}$	75
		Adr1, Spt15, Taf10, Gcn5						
76	2.00	4	rRNA metabolism	$1.50 \cdot 10^{-04}$	75	Nucleolus	$2.87 \cdot 10^{-06}$	100
		Mak5, Nop2, Nop12, Nsa1						
77	2.00	4	GTPase activity	$3.50 \cdot 10^{-04}$	50	No significant terms		
		Sar1, Mf(alpha)1, Cdc7, Tem1						
78	2.00	6	No significant terms			Nucleolus	$6.72 \cdot 10^{-07}$	83
		Nsa2, Nog1, Nog2, Noc2, Ycro723, Mrt4						
79	2.00	4	Small GTPase regulator activity	$3.70 \cdot 10^{-04}$	50	Actin cap	$2.53 \cdot 10^{-05}$	50
		Cla4, Gic1, Rga1, Boi2						

80	2.00	4	Binding	0.014	75	Cytol	0.038	50
		Kap123, Kap104, Sec7, Yap1						
81	2.00	4	No significant terms			Golgi transport complex	$1.44 \cdot 10^{-12}$	100
		Cog2, Cog1, Cog6, Cog4						

Table 13: Statistics for most significant annotation terms (based on GO cellular component) for complete set of modules from Yeast CORE data set

Appendix C

Complete set of modules (Chapter 7)

Module rank	No of Proteins	Cellular Component	P value	Frequency
1	9	snRNP U6	1.56·10-19	78%
		Lsm8, Lsm6, Lsm2, Lsm4, Lsm1, Lsm3, Lsm5, Lsm7, Pat1		
2	9	COPI vesicle coat	1.46·10-16	67%
		Cop1, Ret2, Ret3, Sec21, Sec22, Sec26, Sec27, Bet1, Bos1		
3	13	Pore complex	7.86·10-20	77%
		Nup1, Nup2, Nup42, Nup49, Nup57, Nup60, Nup100, Nup116, Kap93, Kap125, Pab1, Gsp1, Srp1		
4	9	SLIK (SAGA-like) complex	5.59·10-18	78%
		Taf9, Taf1, Taf5, Taf6, Taf10, Gcn4, Ngg1, Ada2, Spt7		
5	8	Oligosaccharyl Transferase Complex	5.39·10-24	100%
		Wbp1, Ost1, Ost2, Ost3, Ost4, Ost5, Swp1, Stt3		
6	8	Proteasome complex	3.67·10-15	88%

		Prc1, Prc2, Pre4, Pre8, Pup3, Scl1, Prn10, Pph22		
7	8	CCR4-NOT complex	5.38·10-23	100%
		Caf40, Caf130, Cdc39, Not3, Not5, Mot2, Pop2, Ccr4		
8	8	Histone deacetylase complex	9.32·10-18	88%
		Hos2, Snt1, Hos4, Set3, Hst1, Zds1, Sif2, Cpr1		
9	7	Pre-replicative complex	1.56·10-19	100%
		Orc1, Orc2, Orc3, Orc4, Orc5, Orc6, Cdc6		
10	10	Nucleolus	3.41·10-9	67%
		Nop2, Nop4, Nop7, Nop15, Nug1, Erb1, Mak21, Nsa2, Sda1, Cka1		
330	8	Nuclear chromatin	1.50·10-6	57%
		Vps1, Nhp10, Arp4, Ceg1, Htb2, Rad16, P02579, Hhf1		
331	6	Eukaryotic 43S preinitiation complex	0.002	33%
		Sui1, Tif5, Ckb2, Myo4, Nip1, Sua7		
332	6	No significant ontology terms	-	-
		Rck1, Ste11, Fus3, Mpt5, Gpa1, Snf1		

333	6	Mating projection tip Cdc42, Bni1, Cla4, Far1, Ste20, Bem1	9.73·10-12	83%
334	7	Incipient bud site Sla2, Tif5, Rpg1, Sui1, Ceg1, Ckb2, Sua7	0.001	29/

Table 14: Statistics for 10 top ranked and 5 bottom ranked modules, based on their most significant annotation terms (GO cellular component) for Yeast CORE data set

Appendix D

Annotation statistics for multi-modular proteins of different module frequency versus single-modular proteins from Yeast CORE data set.

Module frequency	≥1.9	≥1.6	≥1.4	≥1.3	≥1.3	≥1.2	≥1.2	≥1.1	≥1.1	=1
# proteins	50	100	150	200	250	300	350	400	450	450
GO biological process	**GO term frequency** *p* **value**									
GO:0016043 (30%) cellular component organization and biogenesis (30%)	82% $1.3\cdot10^{-11}$	77% $1.1\cdot10^{-19}$	76% $6.2\cdot10^{-29}$	78% $5.4\cdot10^{-42}$	75% $4.5\cdot10^{-47}$	74% $3.2\cdot10^{-54}$	73% $1.0\cdot10^{-60}$	71% $1.5\cdot10^{-66}$	72.7% $4.7\cdot10^{-80}$	66% $2.1\cdot10^{-57}$
GO:0006996 (17.8%) organelle organization and biogenesis (17.8%)	60% $1.2\cdot10^{-08}$	53% $5.8\cdot10^{-13}$	49% $3.0\cdot10^{-16}$	51% $7.4\cdot10^{-24}$	47% $9.3\cdot10^{-25}$	46% $2.9\cdot10^{-27}$	46% $4.9\cdot10^{-33}$	44% $7.2\cdot10^{-34}$	46% $1.3\cdot10^{-43}$	43% $4.5\cdot10^{-36}$
GO:0043283 (30.2%) biopolymer metabolic process (30.2%)	66% $7.2\cdot10^{-05}$	68% $2.8\cdot10^{-12}$	63% $1.8\cdot10^{-14}$	62% $1.3\cdot10^{-17}$	58% $1.4\cdot10^{-17}$	58% $6.0\cdot10^{-21}$	59% $3.5\cdot10^{-28}$	59% $1.2\cdot10^{-31}$	59% $2.4\cdot10^{-36}$	57% $2.3\cdot10^{-30}$
GO:0006139 (20.7%) nucleobase, nucleoside, nucleotide and nucleic acid metabolic process (20.7%)	54% $8.3\cdot10^{-05}$	52% $1.8\cdot10^{-09}$	47% $1.1\cdot10^{-10}$	49% $4.3\cdot10^{-16}$	47% $3.9\cdot10^{-18}$	46% $2.1\cdot10^{-21}$	47% $7.5\cdot10^{-26}$	47% $2.2\cdot10^{-30}$	47% $5.4\cdot10^{-35}$	45% $4.9\cdot10^{-31}$
GO:0016070 (14.2%) metabolic process (14.2%)	42% $5.6\cdot10^{-04}$	40% $8.0\cdot10^{-08}$	37% $2.1\cdot10^{-09}$	39% $3.5\cdot10^{-15}$	36% $9.2\cdot10^{-16}$	36% $6.7\cdot10^{-20}$	37% $6.7\cdot10^{-26}$	37% $2.0\cdot10^{-28}$	37% $3.2\cdot10^{-32}$	34% $8.9\cdot10^{-25}$
GO:0044238 (44.0%) primary metabolic process (44.0%)		74% $4.6\cdot10^{-07}$	70% $4.6\cdot10^{-08}$	69% $8.4\cdot10^{-10}$	68% $4.7\cdot10^{-12}$	68% $2.3\cdot10^{-14}$	69% $1.8\cdot10^{-19}$	69% $8.0\cdot10^{-22}$	68% $3.1\cdot10^{-24}$	67% $5.0\cdot10^{-21}$

Table 15: Statistics for most significant annotation terms of the multi-modular proteins with varying occurrences intervals, compared to the corresponding statistics for single-modular proteins (CORE data set)

Appendix E

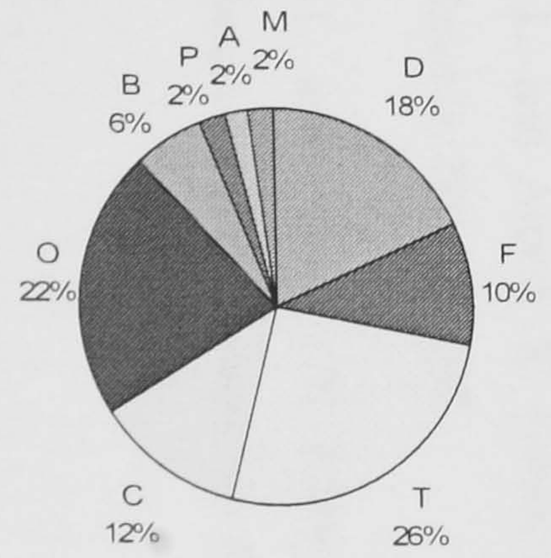
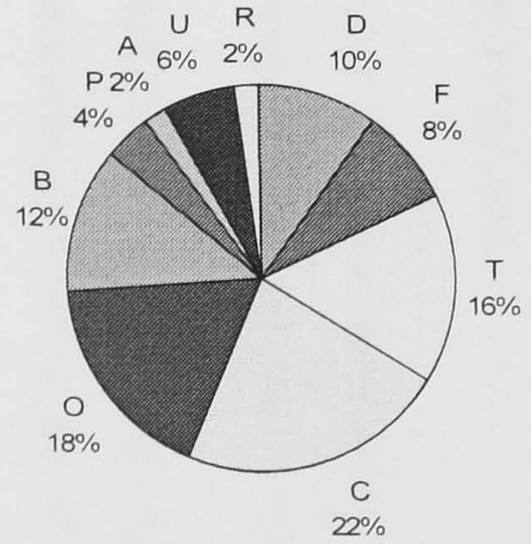
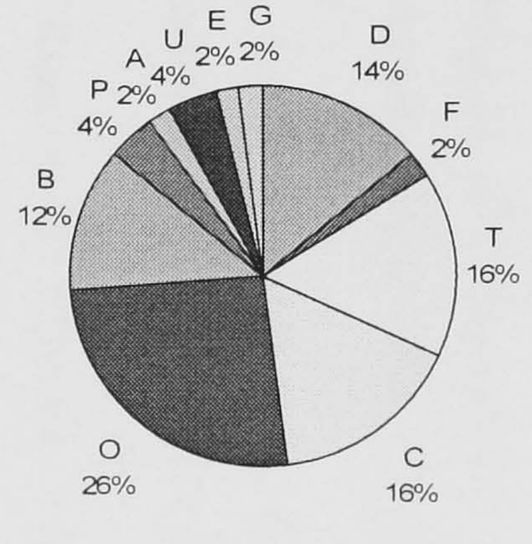
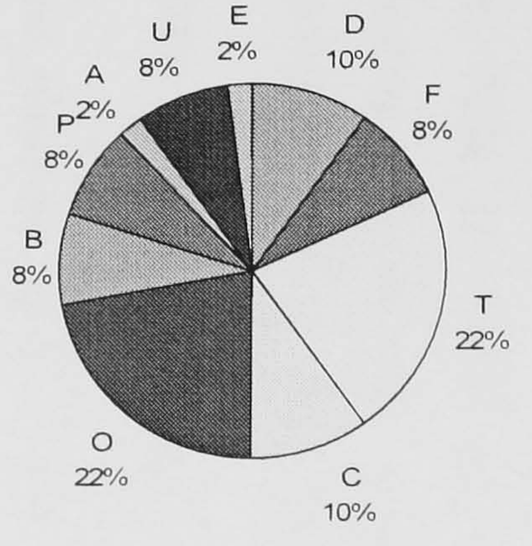
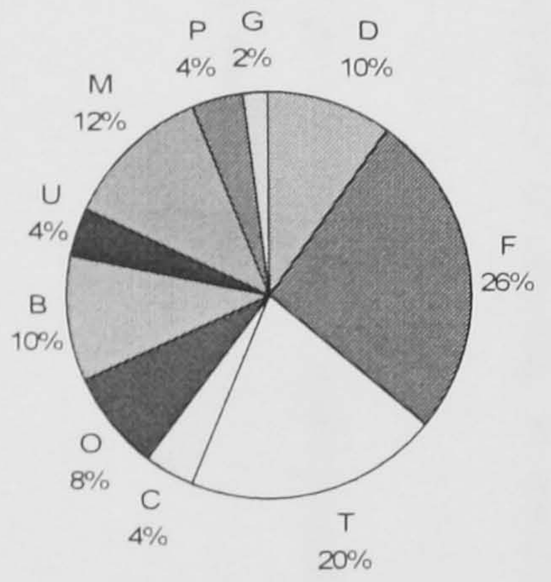
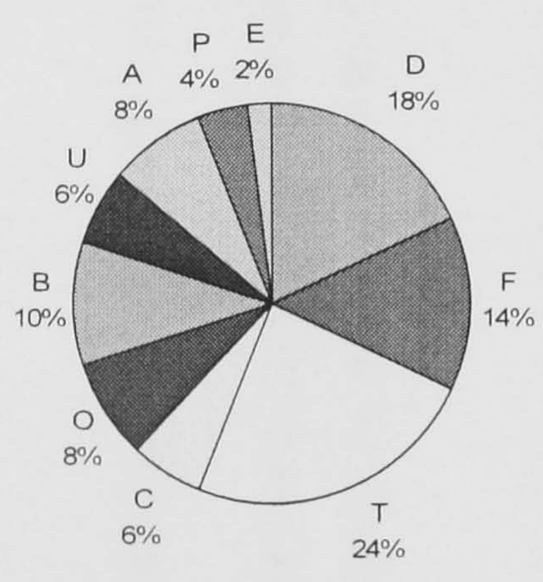
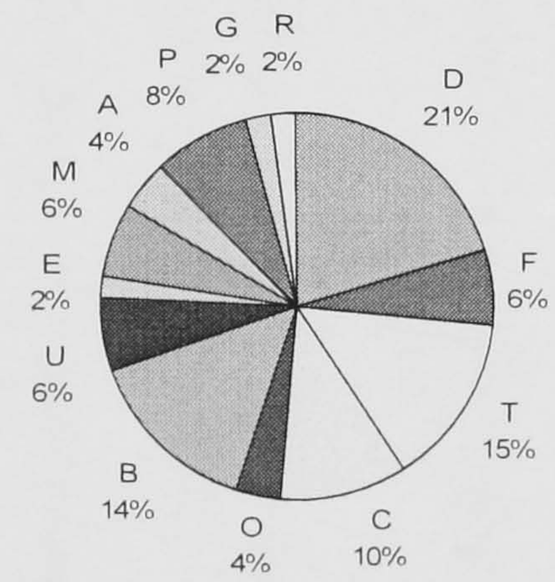
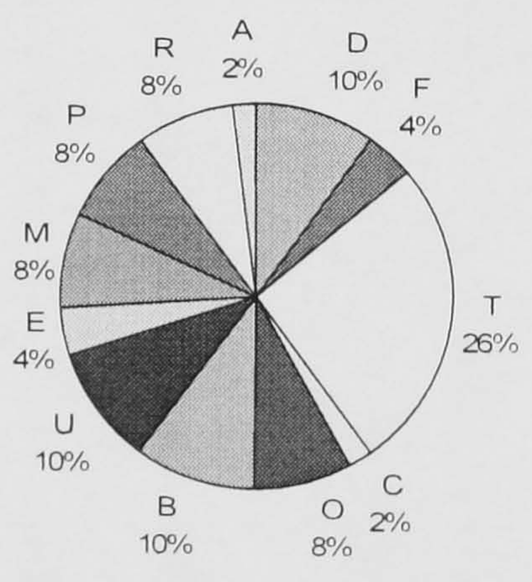
1.9 – 4.2	1.6 – 1.9	1.4 – 1.6	1.3 – 1.4
			
			

Figure 43: Functional groups statistics for proteins in von Mering data set. The first row shows charts with statistics for multi-modular proteins (MMP) in varying intervals of module frequency (in decreasing order of frequency). There are 50 proteins in each interval. For comparison, the second row shows the corresponding statistics for the same number of single-modular proteins

References

(2001) Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425-1433.

Adamcsek, B., Palla, G., Farkas, I.J., Derenyi, I. and Vicsek, T. (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22**, 1021-1023.

Albert, R. (2005) Scale-free networks in cell biology. *J Cell Sci.* **118**, 4947-4957.

Albert, R. and Barabási, A.-L. (2002) Statistical mechanics of complex networks. *Reviews of modern physics*, **74**, 47-97.

Albert, R., Jeong, H. and Barabasi, A.L. (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378-382.

Amaral, L.A., Scala, A., Barthelemy, M. and Stanley, H.E. (2000) Classes of small-world networks. *Proc Natl Acad Sci U S A*, **97**, 11149-11152.

Asthana, S., King, O.D., Gibbons, F.D. and Roth, F.P. (2004) Predicting protein complex membership using probabilistic network reliability. *Genome Res.* **14**, 1170-1175.

Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Bagatelj, V. and Zaversnik, M. (2002) An $O(m)$ Algortihm for Cores Decomposition of Networks, *preprint series*, **40**.

Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509-512.

Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5**, 101-113.

Barrat, A., Barthelemy, M., Pastor-Satorras, R. and Vespignani, A. (2004) The architecture of complex weighted networks. *Proc Natl Acad Sci U S A*, **101**, 3747-3752.

Bernstein, K.A., Gallagher, J.E., Mitchell, B.M., Granneman, S. and Baserga, S.J. (2004) The small-subunit processome is a ribosome assembly intermediate. *Eukaryot Cell*, **3**, 1619-1626.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000) Graph structure in the Web. *Computer Networks: The International Journal of Computer and Telecommunications Networking* **33**.

Cohen, R. and Havlin, S. (2003) Scale-free networks are ultrasmall. *Phys Rev Lett*, **90**, 058701.

Conant, G.C. and Wagner, A. (2003) Convergent evolution of gene circuits. *Nat Genet*, **34**, 264-266.

Cullen, P.J. and Sprague, G.F., Jr. (2000) Glucose depletion causes haploid invasive growth in yeast. *Proc Natl Acad Sci U S A*, **97**, 13619-13624.

Deng, M., Tu, Z., Sun, F. and Chen, T. (2004) Mapping Gene Ontology to proteins based on protein-protein interaction data, *Bioinformatics*, **20**, 895-902.

Dwight, S.S., Harris, M.A., Dolinski, K., Ball, C.A., Binkley, G., Christie, K.R., Fisk, D.G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D. and Cherry, J.M. (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res*, **30**, 69-72.

Edgington, N.P., Blacketer, M.J., Bierwagen, T.A. and Myers, A.M. (1999) Control of Saccharomyces cerevisiae filamentous growth by cyclin-dependent kinase Cdc28. *Mol Cell Biol*, **19**, 1369-1380.

Eisenberg, D., Marcotte, E.M., Xenarios, I. and Yeates, T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823-826.

Enright, A.J., Van Dongen, S. and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584.

Erdős, P. and Rényi, A. (1959) On random graphs. *Publicationes Mathematicae*. **6**, 290-297.

Fields, S. and Bartel, P.L. (2001) The two-hybrid system. A personal view. *Methods Mol Biol.* **177**, 3-8.

Flake, G.W., Lawrence, C., Giles, C.L. and Coetzee, F.M. (2002) Self-organization and identification of Web communities, *IEEE Computer*. **35**, 66-71.

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. and Feldman, M.W. (2002) Evolutionary rate in the protein interaction network. *Science*. **296**, 750-752.

Freeman, L.C. (1979) Centrality in Social Networks: Conceptual Clarification. *Social networks*. **1**, 215-239.

Fromont-Racine, M., Mayes, A.E., Brunet-Simon, A., Rain, J.C., Colley, A., Dix, I., Decourty, L., Joly, N., Ricard, F., Beggs, J.D. and Legrain, P. (2000) Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins, *Yeast*. **17**, 95-110.

Gagneur, J., Krause, R., Bouwmeester, T. and Casari, G. (2004) Modular decomposition of protein-protein interaction networks. *Genome Biol.* **5**, R57.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B. and Superti-Furga, G. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*. **440**, 631-636.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C.,

Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, **415**, 141-147.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley, R.L., Jr., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J. and Rothberg, J.M. (2003) A protein interaction map of Drosophila melanogaster, *Science*, **302**, 1727-1736.

Girvan, M. and Newman, M.E. (2002) Community structure in social and biological networks, *Proc Natl Acad Sci U S A*, **99**, 7821-7826.

Goldberg, D.S. and Roth, F.P. (2003) Assessing experimentally derived interactions in a small world, *Proc Natl Acad Sci U S A*, **100**, 4372-4376.

Guglielmi, B., van Berkum, N.L., Klapholz, B., Bijma, T., Boube, M., Boschiero, C., Bourbon, H.M., Holstege, F.C. and Werner, M. (2004) A high resolution protein interaction map of the yeast Mediator complex, *Nucleic Acids Res*, **32**, 5379-5391.

Guldener, U., Munsterkotter, M., Kastenmuller, G., Strack, N., van Helden, J., Lemer, C., Richelles, J., Wodak, S.J., Garcia-Martinez, J., Perez-Ortin, J.E., Michael, H., Kaps, A., Talla, E., Dujon, B., Andre, B., Souciet, J.L., De Montigny, J., Bon, E., Gaillardin, C. and Mewes, H.W. (2005) CYGD: the Comprehensive Yeast Genome Database, *Nucleic Acids Res*, **33**, D364-368.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P. and Vidal, M. (2004) Evidence for

dynamically organized modularity in the yeast protein-protein interaction network. *Nature*. **430**, 88-93.

Hartuv, E. and Shamir, R. (2000) A clustering algorithm based on graph connectivity. *Information Processing Letters*. **76**, 175-181.

Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47-52.

He, W. and Parker, R. (2000) Functions of Lsm proteins in mRNA degradation and splicing, *Curr Opin Cell Biol*. **12**, 346-350.

Hoglund, M., Frigyesi, A. and Mitelman, F. (2006) A gene fusion network in human neoplasia. *Oncogene*. **25**, 2674-2678.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002) Revealing modular organization in the yeast transcriptional network, *Nat Genet*, **31**, 370-377.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, **98**, 4569-4574.

Ito, T., Ota, K., Kubota, H., Yamaguchi, Y., Chiba, T., Sakuraba, K. and Yoshida, M. (2002) Roles for the two-hybrid system in exploration of the yeast protein interactome, *Mol Cell Proteomics*. **1**, 561-566.

Jansen, R., Lan, N., Qian, J. and Gerstein, M. (2002) Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics*. **2**, 71-81.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*. **302**, 449-453.

Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature*. **411**, 41-42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) The large-scale organization of metabolic networks, *Nature*, **407**, 651-654.

Jiang, J.J. and Conrath, D.W. (1998) Semantic similarity based on corpus statistics and lexical taxonomy. *International Conference on Research in Computational Linguistics*, Taiwan, 19-33.

Kao, L.R., Peterson, J., Ji, R., Bender, L. and Bender, A. (1996) Interactions between the ankyrin repeat-containing protein Akr1p and the pheromone response pathway in Saccharomyces cerevisiae, *Mol Cell Biol*, **16**, 168-178.

Karaoz, U., Murali, T.M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C.R. and Kasif, S. (2004) Whole-genome annotation by using evidence integration in functional-linkage networks, *Proc Natl Acad Sci U S A*, **101**, 2888-2893.

Kauffman, S.A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets, *J Theor Biol*, **22**, 437-467.

Klingenhoff, A., Frech, K. and Werner, T. (2002) Regulatory modules shared within gene classes as well as across gene classes can be detected by the same in silico approach, *In Silico Biol*, **2**, S17-26.

Knauer, R. and Lehle, L. (1999) The oligosaccharyltransferase complex from Saccharomyces cerevisiae. Isolation of the OST6 gene, its synthetic interaction with OST3, and analysis of the native complex, *J Biol Chem*, **274**, 17249-17256.

Kozminski, K.G., Beven, L., Angerman, E., Tong, A.H., Boone, C. and Park, H.O. (2003) Interaction between a Ras and a Rho GTPase couples selection of a growth site to the development of cell polarity in yeast, *Mol Biol Cell*, **14**, 4958-4970.

Lee, P.R., Song, S., Ro, H.S., Park, C.J., Lippincott, J., Li, R., Pringle, J.R., De Virgilio, C., Longtine, M.S. and Lee, K.S. (2002) Bni5p, a septin-interacting protein, is required for normal septin function and cytokinesis in Saccharomyces cerevisiae, *Mol Cell Biol*, **22**, 6906-6920.

Lengeler, K.B., Davidson, R.C., D'Souza, C., Harashima, T., Shen, W.C., Wang, P., Pan, X., Waugh, M. and Heitman, J. (2000) Signal transduction cascades regulating fungal development and virulence. *Microbiol Mol Biol Rev*, **64**, 746-785.

Lin, D. (1998) An information-theoretic definition of similarity. *The 15th International Conference on Mashine Learning*. Madison, WI. 296-304.

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275-1283.

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A. (2003) Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput*, 601-612.

Lubovac, Z., Corne, D., Gamalielsson, J. and Olsson, B. (2007) Weighted cohesiveness for identification of functional modules and their interconnectivity. In Hochreiter, S. and Wagner, R. (eds), *Proceedings of Bioinformatics in Research and Development (BIRD 2007)*. LNBI, Berlin, 185-198.

Lubovac, Z., Gamalielsson, J. and Olsson, B. (2005) Combining topological characteristics and domain knowledge reveals functional modules in protien interaction networks. In Sagot, M.-F. and Gumarães, K.S. (eds), *Proceedings of CompBioNets - Algorithms and Computational Methods for Biochemical and Evolutionary Networks*. College Publications, Lyon, France, 93-106.

Lubovac, Z., Gamalielsson, J. and Olsson, B. (2006) Combining functional and topological properties to identify core modules in protein interaction networks, *Proteins*, **64**, 948-959.

Lubovac, Z., Gamalielsson, J., Olsson, B. and Lindlöf, A. (2005) Exploring protein networks with a semantic similarity measure across Gene Ontology. *Proceedings of the 6th International Symposium on Computational Biology and Genome Informatics*. Salt Lake City, USA, 1203-1208.

Luo, F. and Scheuermann, R.H. (2006) Detecting Functional Modules from Protein Interaction Networks. *Proceeding of the First International Multi-Symposiums on Computer and Computational Scineces (IMSCCS'06)*. IEEE Computer Society.

Luo, F., Yang, Y., Chen, C.F., Chang, R., Zhou, J. and Scheuermann, R.H. (2007) Modular organization of protein interaction networks, *Bioinformatics*, **23**, 207-214.

Mangan, S., Zaslaver, A. and Alon, U. (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks, *J Mol Biol*, **334**, 197-204.

Mann, M., Hendrickson, R.C. and Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry, *Annu Rev Biochem*, **70**, 437-473.

Maslov, S. and Sneppen, K. (2002) Specificity and stability in topology of protein networks, *Science*, **296**, 910-913.

Mendenhall, M.D. and Hodge, A.E. (1998) Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast Saccharomyces cerevisiae, *Microbiol Mol Biol Rev*, **62**, 1191-1243.

Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B. (2002) MIPS: a database for genomes and protein sequences, *Nucleic Acids Res*, **30**, 31-34.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) Network motifs: simple building blocks of complex networks, *Science*, **298**, 824-827.

Mori, Y., Nishii, H., Takabe, K., Shinozaki, H., Matsumoto, N., Suzuki, K., Tanabe, H., Watanabe, A., Ochiai, K. and Tanaka, T. (2003) Preoperative diagnosis of malignant transformation arising from mature cystic teratoma of the ovary, *Gynecol Oncol*, **90**, 338-341.

Mosch, H.U., Kubler, E., Krappmann, S., Fink, G.R. and Braus, G.H. (1999) Crosstalk between the Ras2p-controlled mitogen-activated protein kinase and cAMP

pathways during invasive growth of Saccharomyces cerevisiae. *Mol Biol Cell*. **10**. 1325-1335.

Mosch. H.U., Roberts, R.L. and Fink, G.R. (1996) Ras2 signals via the Cdc42/Ste20/mitogen-activated protein kinase module to induce filamentous growth in Saccharomyces cerevisiae, *Proc Natl Acad Sci U S A*. **93**. 5352-5356.

Mrowka, R., Patzak, A. and Herzel, H. (2001) Is there a bias in proteome research?. *Genome Res*. **11**, 1971-1973.

Newman, M.E.J. (2001) The structure of scientific collaboration networks. *Proc Natl Acad Sci U S A*, **98**, 404-409.

Newman, M.E.J. (2002) Assortative mixing in networks, *Phys Rev Lett*, **89**, 208701.

Newman, M.E.J. (2003) Ego-centered networks and the ripple effect. *Socal networks*. **25**, 83-95.

Newman, M.E.J. (2003) The structure and function of complex networks. *SIAM Reviews*. **45**. 167-256.

Newman, M.E.J. (2004) Detecting community structure in networks, *European Physical Journal B*, **38**, 321-330.

Oltvai, Z.N. and Barabasi, A.L. (2002) Systems biology. Life's complexity pyramid, *Science*, **298**, 763-764.

Obermann, E.C., Went, P., Zimpfer, A., Tzankov, A., Wild, P.J., Stoehr, R., Pileri, S.A. and Dirnhofer, S. (2005) Expression of minichromosome maintenance protein 2 as a marker for proliferation and prognosis in diffuse large B-cell lymphoma: a tissue microarray and clinico-pathological analysis. *BMC Cancer*. **5**. 162.

Onnela, J.P., Saramaki, J., Kertesz, J. and Kaski, K. (2005) Intensity and coherence of motifs in weighted complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*. **71**. 065103.

Pan, X. and Heitman, J. (1999) Cyclic AMP-dependent protein kinase regulates pseudohyphal differentiation in Saccharomyces cerevisiae. *Mol Cell Biol.* **19**, 4874-4887.

Pereira-Leal, J.B., Enright, A.J. and Ouzounis, C.A. (2004) Detection of functional modules from protein interaction networks, *Proteins*, **54**, 49-57.

Pereira-Leal, J.B., Levy, E.D. and Teichmann, S.A. (2006) The origins and evolution of functional modules: lessons from protein complexes. *Philos Trans R Soc Lond B Biol Sci*, **361**. 507-517.

Petti, A.A. and Church, G.M. (2005) A network of transcriptionally coordinated functional modules in Saccharomyces cerevisiae. *Genome Res.* **15**. 1298-1306.

Peulen, O., Dewé, W., Dandrifosse, G., Henrotay, I. and Romain, N. (1998) The relationship between spermine content of human milk during the first postnatal month and allergy in children, *Public Health Nutrition*, **1**, 181-184.

Posas, F. and Saito, H. (1997) Osmotic activation of the HOG MAPK pathway via Ste11p MAPKKK: scaffold role of Pbs2p MAPKK. *Science*. **276**. 1702-1705.

Posse, C., Sanfilippo, A., Goplan, B., Riensche, R., Beagley, N. and Baddeley, B. (2006) Cross-Ontological Analytics:Combining associative and hierarchical relations in the Gene Ontologies to assess gene product similarity., *Proceedings of Interational Workshop on Bioinformatics Reasearch and Applications*., Reading, UK.

Poyatos, J.F. and Hurst, L.D. (2004) How biologically relevant are interaction-based modules in protein networks?, *Genome Biol.* **5**, R93.

Prinz, S., Avila-Campillo, I., Aldridge, C., Srinivasan, A., Dimitrov, K., Siegel, A.F. and Galitski, T. (2004) Control of yeast filamentous-form growth by modules in an integrated molecular network. *Genome Res.* **14**. 380-390.

Przulj, N., Wigle, D.A. and Jurisica, I. (2004) Functional topology in a network of protein interactions, *Bioinformatics*, **20**, 340-348.

Qi, Y. and Ge, H. (2006) Modularity and dynamics of cellular networks. *PLoS Comput Biol*, **2**, e174.

Rader, S.D. and Guthrie, C. (2002) A conserved Lsm-interaction motif in Prp24 required for efficient U4/U6 di-snRNP formation. *Rna*, **8**, 1378-1392.

Ravasz, E. and Barabasi, A.L. (2003) Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, **67**, 026112.

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. and Barabasi, A.L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551-1555.

Redner, S. (1998) How popular is youi paper? An empirical study of the citation distribution, *European Physical Journal B*, **4**, 131-134.

Resnik, P. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, **11**, 95-130.

Rives, A.W. and Galitski, T. (2003) Modular organization of cellular networks. *Proc Natl Acad Sci U S A*, **100**, 1128-1133.

Rosen, K.H. (1995) *Discrete mathematics and its applications*. McGraw-Hill.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M. and Mewes, H.W. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*, **32**, 5539-5545.

Rupp, S., Summers, E., Lo, H.J., Madhani, H. and Fink, G. (1999) MAP kinase and cAMP filamentation signaling pathways converge on the unusually large promoter of the yeast FLO11 gene. *Embo J*, **18**, 1257-1269.

Salwinski, L. and Eisenberg, D. (2003) Computational methods of analysis of protein-protein interactions. *Curr Opin Struct Biol*, **13**, 377-382.

Schlosser (2004) The role of modules in development and evolution. In Schlosser. G. and Wagner. G.P. (eds). *Modularity in development and evolution*. The University of Chicago Press. 519-582.

Schlosser, G. and Wagner. G.P. (2004) *Modularity in development and evolution*. The University of Chicago Press.

Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein-protein interactions in yeast, *Nat Biotechnol*, **18**. 1257-1261.

Segal, E., Friedman, N., Koller, D. and Regev. A. (2004) A module map showing conditional activity of expression modules in cancer. *Nat Genet*. **36**. 1090-1098.

Segal, E., Shapira, M., Regev. A., Pe'er. D., Botstein. D., Koller. D. and Friedman. N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*. **34**. 166-176.

Seo, J. and Schneiderman, B. (2002) Interactively Exploring Hierarchical Clustering Results, *IEEE Computer*. **35**. 80-86.

Spirin, V. and Mirny. L.A. (2003) Protein complexes and functional modules in molecular networks, *Proc Natl Acad Sci U S A*. **100**, 12123-12128.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. and Gilles, E.D. (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190-193.

Stevens, S.W. and Abelson, J. (1999) Purification of the yeast U4/U6.U5 small nuclear ribonucleoprotein particle and identification of its proteins. *Proc Natl Acad Sci U S A*. **96**. 7226-7231.

Strogatz. S.H. (2001) Exploring complex networks. *Nature*. **410**. 268-276.

Takeuchi, J. and Tamura, T. (2004) Recombinant ATPases of the yeast 26S proteasome activate protein degradation by the 20S proteasome. *FEBS Lett*. **565**. 39-42.

Tang, H.Y. and Cai, M. (1996) The EH-domain-containing protein Pan1 is required for normal organization of the actin cytoskeleton in Saccharomyces cerevisiae. *Mol Cell Biol*. **16**, 4897-4914.

Tape, T.G. (2005) Interpreting diagnostic tests.

Tharun, S., He, W., Mayes, A.E., Lennertz, P., Beggs, J.D. and Parker, R. (2000) Yeast Sm-like proteins function in mRNA decapping and decay. *Nature*, **404**, 515-518.

Titz, B., Schlesner, M. and Uetz, P. (2004) What do we learn from high-throughput protein interaction data?. *Expert Rev Proteomics*, **1**, 111-121.

Tong, A.H., Drees, B., Nardelli, G., Bader, G.D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., Quondam, M., Zucconi, A., Hogue, C.W., Fields, S., Boone, C. and Cesareni, G. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science*, **295**, 321-324.

Tornow, S. and Mewes, H.W. (2003) Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res*, **31**, 6283-6289.

Uhrig, J.F. (2006) Protein interaction networks in plants. *Planta*, **224**, 771-781.

Van Criekinge, W. and Beyaert, R. (1999) Yeast Two-Hybrid: State of the Art. *Biol Proced Online*, **2**, 1-38.

van Noort, V., Snel, B. and Huynen, M.A. (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep*, **5**, 280-284.

Watts, D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440-442.

Versele, M. and Thorner, J. (2004) Septin collar formation in budding yeast requires GTP binding and direct phosphorylation by the PAK, Cla4. *J Cell Biol*, **164**, 701-715.

Winters, M.S. and Day, R.A. (2003) Detecting protein-protein interactions in the intact cell of Bacillus subtilis (ATCC 6633). *J Bacteriol*, **185**, 4268-4275.

Vogelstein, B., Lane, D. and Levine, A.J. (2000) Surfing the p53 network. *Nature*, **408**, 307-310.

Wollenberg, K. and Swaffield, J.C. (2001) Evolution of proteasomal ATPases. *Mol Biol Evol*, **18**, 962-974.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions, *Nature*, **417**, 399-403.

Wuchty, S. and Almaas, E. (2005) Peeling the yeast protein network, *Proteomics*, **5**, 444-449.

Wuchty, S., Oltvai, Z.N. and Barabasi, A.L. (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network, *Nat Genet*, **35**, 176-179.

Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M. and Eisenberg, D. (2000) DIP: the database of interacting proteins, *Nucleic Acids Res*, **28**, 289-291.

Yook, S.H., Jeong, H., Barabasi, A.L. and Tu, Y. (2001) Weighted evolving networks, *Phys Rev Lett*, **86**, 5835-5838.

Yook, S.H., Oltvai, Z.N. and Barabasi, A.L. (2004) Functional and topological characterization of protein interaction networks, *Proteomics*, **4**, 928-942.

Yuan, Y.O., Stroke, I.L. and Fields, S. (1993) Coupling of cell identity to signal response in yeast: interaction between the alpha 1 and STE12 proteins, *Genes Dev*, **7**, 1584-1597.

Zachariae, W., Shin, T.H., Galova, M., Obermaier, B. and Nasmyth, K. (1996) Identification of subunits of the anaphase-promoting complex of Saccharomyces cerevisiae, *Science*, **274**, 1201-1204.